Extended Hypervolume Improvement Distributions for Constrained Multi-Objective Bayesian Optimization with Correlated Gaussian Processes

Ryan Dorland (s3219992)

Supervisors:
Dr. H. Wang & Dr. E. Raponi

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

June 5, 2026

Abstract

Bayesian optimization is the standard approach for expensive multi-objective problems, and recent work by Wang et al. has shown that acquisition functions built on the full probability distribution of the hypervolume improvement (HVI), rather than on a single moment of it, can substantially accelerate convergence. Their derivation, however, rests on two assumptions that limit its reach: the two objective Gaussian-process posteriors must be independent, and no inequality constraints are admitted. This thesis closes both gaps. First, the cell-partition machinery is extended to a bivariate Gaussian posterior with arbitrary correlation, replacing the product of univariate CDFs by a bivariate-normal rectangle probability and revealing a Laurent-polynomial structure in the within-cell density; a consistency theorem confirms exact reduction to the independent baseline at $\rho = 0$. Second, the constrained HVI distribution is derived under both independent and correlated objective-constraint posteriors, yielding a constrained ε-PoHVI acquisition function that handles the induced dependence between objectives created by conditioning on a shared constraint. This constrained extension is verified numerically against Monte Carlo simulation but left for future end-to-end BO benchmarking. Both derivations are verified against 100,000-sample Monte Carlo simulation across 526 configurations including 200 ICM-fitted posteriors, agreeing within the expected empirical-CDF sampling error. A multi-output GP benchmark on the WOSGZ family selects an Intrinsic Coregionalization Model surrogate, and a full Bayesian-optimization comparison on 13 standard problems shows that the correlated pipeline wins on every problem with exploitable inter-objective coupling, all eight WOSGZ problems (15/15 seeds, +28% to +100% relative HV gains) plus ZDT6 (+53%), and on the remaining four ZDT problems ties on three (ZDT1, ZDT3, ZDT4) and loses on one (ZDT2 at $p_{\text{Holm}} = 0.0085$, -3.6%). Per-seed variance is also reduced by 2-8$\times$ on the WOSGZ family. The empirical comparison varies surrogate (independent GPs vs. ICM) and acquisition (independent HVI vs. correlated HVI) jointly; isolating the contribution of the correlated HVI distribution alone is left to future work. As indirect evidence consistent with the correlated distribution contributing to the gain, the BO-trajectory $|\bar{\rho}|$ values are non-trivial on every problem the correlated pipeline wins.

Contents

1 Introduction

1.1 The Problem: Expensive Multi-Objective Optimization Under Constraints

Many expensive design problems require optimizing several competing objectives subject to hard constraints. Consider a materials engineer choosing a microstructure for a load-bearing component: compliance (deflection under load) and density are physically coupled through the underlying microstructure, meaning any change to the material shifts both at once. Furthermore, the design must keep peak stress under load below a yield limit. The objectives are physically coupled, the constraint shares variables with the objectives, and every candidate is expensive to evaluate because it requires either a finite-element simulation or a fabricated test piece. This kind of physical coupling, between objectives or between objectives and constraints, is exactly the structure this thesis sets out to exploit.

Two features distinguish this regime from textbook optimization. First, evaluations are costly: a physical experiment, a multi-hour simulation, or a lengthy training run. Second, no closed-form expressions or gradients are available. The objectives and constraints are black boxes. Together they create a strong incentive to choose each new evaluation point carefully, and to use the full posterior, not just its mean, when deciding where to evaluate next.

Bayesian optimization (BO) [26] addresses this setting by constructing a probabilistic surrogate model, typically a Gaussian process (GP), from observed data and using an *acquisition function* to select the next point to evaluate. The acquisition function quantifies the expected benefit of evaluating a candidate, balancing exploitation of regions that look promising against exploration of regions where the model is uncertain. In the single-objective case, acquisition functions such as Expected Improvement (EI) and Probability of Improvement (PoI) have a long history and well-understood theoretical properties [19, 22].

When multiple objectives are present, the notion of a single optimum is replaced by the Pareto front: the set of solutions for which no objective can be improved without degrading another. The hypervolume indicator (HV) [32] measures the quality of a finite approximation to the Pareto front by computing the volume of objective space that is dominated by the approximation set and bounded by a reference point. The hypervolume improvement (HVI) quantifies the gain in this indicator when a new point is added, and serves as the foundation for several successful multi-objective acquisition functions, including Expected Hypervolume Improvement (EHVI) [12, 30] and Probability of Improvement (PoI) [12].

1.2 From Moments to Distributions

A limitation shared by EHVI and standard PoI is that they each use only a single summary statistic of the HVI random variable: EHVI uses the mean, and PoI uses the probability of exceeding zero. Two candidates with identical EHVI can have very different probabilities of producing a practically meaningful improvement; conversely, a high-EHVI candidate may place much of its probability mass on negligible or zero improvement. A distributional acquisition can distinguish these cases, a moment-based one cannot.

This observation motivated Wang et al. [27] to derive the probability distribution of HVI under a bivariate Gaussian posterior. Having access to the full distribution, rather than just a single

moment, enables a family of acquisition functions indexed by a threshold parameter. Wang et al.'s proposed ε-Probability of Hypervolume Improvement (ε-PoHVI) computes

$$\varepsilon\text{-PoHVI}(x) = \Pr\big(\Delta(y) > \varepsilon \mid \mathcal{D}, x\big), \tag{1}$$

the probability that evaluating a candidate point x yields at least ε hypervolume improvement over the current approximation set. By varying ε, the practitioner controls the quality threshold demanded of new evaluations: large ε focuses search on high-impact regions, while small ε encourages broader exploration. Experimentally, ε-PoHVI showed substantial convergence improvements over EHVI on problems with irregular Pareto fronts and high GP uncertainty [27]. The derivation relies on a cell-partition framework introduced in Section 2.5.

1.3 Two Open Limitations

The framework of Wang et al. [27] rests on two assumptions that limit its applicability to realistic problems, both identified by the authors as future work [27, Section 6].

Independent objective posteriors. The derivation assumes that the GP posteriors for the two objectives are independent: $y_1 \perp y_2 \mid \mathcal{D}$. This assumption allows cell probabilities to factor as products of univariate CDF differences and it allows the integrand of the product density to separate into a product of two univariate Gaussian densities. Yet in practice, objectives are often statistically dependent. Multi-output GP models such as the linear model of coregionalization [2], convolution processes [7, 1], and intrinsic coregionalization models [6] are designed to capture cross-correlations between outputs, because modeling these dependencies improves predictive accuracy and data efficiency. When such models are used, the posterior at any candidate point generically has non-zero correlation $\rho = \mathrm{Corr}(y_1, y_2) \neq 0$, and the independent-objective HVI distribution no longer applies.

No constraint handling. The framework considers only unconstrained problems. However, many real-world optimization tasks involve inequality constraints that must be satisfied: a design is useless if it violates a stress limit, regardless of how well it performs on the objectives. The standard approach to handling constraints in Bayesian optimization, due to Gardner et al. [14], is the *product weighting*: multiply the acquisition function by the posterior probability of feasibility,

$$\alpha_{\text{constrained}}(x) = \alpha(x) \cdot \prod_{j=1}^{J} \Pr\big(c_j(x) \leq 0 \mid \mathcal{D}\big). \tag{2}$$

This formula treats feasibility and objective quality as independent factors. When the constraint GPs are independent of the objective GPs, that treatment is adequate. But when constraints and objectives are modeled jointly, as occurs naturally when a multi-output GP is used or when a constraint is physically related to an objective, conditioning on feasibility changes the distribution of the objectives. The product formula then uses the wrong (unconditional) HVI distribution, and the acquisition values it assigns can be systematically incorrect.

Shared Underlying Mechanism The two limitations share a single underlying mechanism. Whenever the bivariate posterior at a candidate has a non-trivial off-diagonal covariance, whether that covariance is produced by a multi-output GP surrogate (the correlation case) or induced by conditioning on a correlated constraint (the constraint case), the two objectives at that candidate are dependent. The Wang et al. framework, which factors objective behavior into a product of two univariate Gaussians, cannot represent that dependence. Resolving either limitation therefore requires the same core technical building block: a bivariate truncated-Gaussian treatment of the cell integrals. Though, the constrained case adds further machinery on top: a mixture decomposition by feasibility and, for multiple correlated constraints, multivariate orthant probabilities. The two contributions of this thesis are consequences of that single observation, not parallel fixes to two separate problems.

To our knowledge, the Wang et al. full-distribution framework has not been extended to correlated bivariate Gaussian objective posteriors, nor to constraints incorporated directly into the HVI random variable. Existing constrained BO methods [14, 15, 21, 23, 8] apply feasibility weighting or optimize expected quantities; they do not provide the full constrained HVI distribution required by ε-PoHVI-style threshold acquisitions.

1.4 Research Question

The central question of this thesis is:

> *How can the full probability distribution of generalized hypervolume improvement be extended from independent bivariate Gaussian posteriors to correlated objective posteriors and to inequality-constrained settings?*

This decomposes into four sub-questions:

1. How must the cell-partition methodology of Wang et al. [27] change when the two objective posteriors form a correlated bivariate Gaussian?
2. How can feasibility be incorporated directly into the HVI random variable, rather than applied as an external product weight?
3. Can the resulting correlated and constrained distributions be evaluated accurately enough for use inside acquisition functions, and do they agree with Monte Carlo simulation?
4. Does the correlated-objective acquisition function improve empirical convergence over the independent ε-PoHVI baseline on benchmark problems with exploitable cross-objective structure?

1.5 Contributions

This thesis makes two theoretical contributions and one empirical validation contribution.

Contribution 1: The full HVI distribution becomes available for any multi-output GP surrogate (Section 3). This thesis derives the probability density and cumulative distribution function of the generalized hypervolume improvement when the bivariate posterior carries correlation $\rho \in (-1, 1)$. The independent cell probabilities of Wang et al. are replaced by bivariate normal rectangle probabilities; the within-cell product density retains a Laurent-polynomial exponent that preserves the one-dimensional quadrature recipe of the independent case. A consistency theorem (Theorem 3.20) verifies that all formulas reduce exactly to the Wang et al. results when $\rho = 0$.

Contribution 2: Feasibility can be incorporated directly into the HVI random variable (Section 4). The constrained generalized HVI random variable $\Delta_C = \Delta(y) \cdot \prod_j \mathbf{1}[g_j \leq 0]$ is defined, and its distribution is derived under independent constraints, a single correlated constraint, and multiple correlated constraints. The distribution takes the form of a mixture of an infeasible point mass at zero and a feasible continuous component. Under correlated constraints the feasible component requires Contribution 1 as a subroutine, because conditioning on a shared constraint induces dependence between the objectives even when their unconditional posterior is independent. For a single correlated constraint the result reduces to a one-dimensional outer integral over the truncated Gaussian constraint distribution; for multiple correlated constraints the evaluation requires multivariate Gaussian orthant probabilities, computed via quasi-Monte Carlo. The result is a constrained acquisition function ε-CPoHVI that computes the probability of *feasible* hypervolume improvement of size at least ε. *Scope:* this contribution is derived and numerically verified in this thesis; a full constrained BO comparison against the standard constrained MOBO baselines, for example the constrained expected hypervolume improvement of Feliot et al. [13], the noisy parallel variants NEHVI-C / qNEHVI-C [8], the entropy-based MESMOC [4], and PF^{2}ES [24], is left to future work.

Contribution 3: Empirical validation and end-to-end benchmark (Sections 6 to 8). The analytical distributions are verified against 10^5-sample Monte Carlo simulation across 526 configurations and agree within the expected empirical-CDF sampling error, including 200 realistic ICM-fitted posteriors drawn from ZDT1, ZDT2, and the full WOSGZ1 to WOSGZ8 family. Four multi-output GP architectures are benchmarked against an independent-GP baseline on the WOSGZ family and GPyTorch ICM rank 1 is selected; with this surrogate, corr-ε-PoHVI is compared against the Wang et al. baseline on 13 bi-objective problems with paired-by-seed methodology and Holm-Bonferroni correction across the 13-problem family. The benchmark focuses on the correlated-objective acquisition; ε-CPoHVI is left for the future-work benchmark identified above.

1.6 Thesis Outline

The remainder of this thesis is organized as follows. Section 2 (Preliminaries) collects the definitions, notation, and foundational results used throughout the thesis. Section 3 (Extension I: Correlated Gaussian Objectives) develops Contribution 1. Section 4 (Extension II: Constrained Hypervolume Improvement) defines the constrained HVI random variable, derives its distribution under independent constraints, a single correlated constraint, and multiple correlated constraints, and defines the constrained ε-PoHVI acquisition function. Section 5 discusses what correlation and constraints

change about the HVI distribution and how the two contributions relate. Section 6 (Acquisition Functions and Numerical Verification) turns the derivations into executable acquisitions and verifies them numerically. Section 7 (Surrogate Model Selection) selects the multi-output GP for the correlated acquisition. Section 8 (Bayesian Optimization Comparison) reports the 13-problem benchmark. Section 9 summarizes the empirical findings and discusses limitations and directions for future work.

2 Preliminaries

This chapter collects the definitions, notation, and foundational results that are used throughout the thesis. This includes all standard Gaussian identities, the hypervolume indicator framework, the cell partition machinery, and the product-of-variables density lemma.

2.1 Gaussian Process Posteriors

Each objective function is modeled as the realization of a centered Gaussian process (GP) prior [25]:

$$f_i \sim \mathcal{GP}(0,\, k_i), \qquad i \in [1..m], \tag{3}$$

where $k_i \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel function modeling the auto-covariance of f_i, so that $\mathrm{Cov}\{f_i(\mathbf{x}),\, f_i(\mathbf{x}')\} =_i (\mathbf{x}, \mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

Given a data set $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ with $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$ and $\mathbf{Y} = \{\mathbf{f}(\mathbf{x}^{(1)}), \ldots, \mathbf{f}(\mathbf{x}^{(n)})\}$, the posterior GPs are $f_i \mid \mathcal{D} \sim \mathcal{GP}(\hat{f}_i,\, \hat{k}_i)$, where $\hat{f}_i$ and $\hat{k}_i$ denote the posterior mean and kernel functions, respectively. At an unevaluated candidate point $\mathbf{x} \in \mathcal{X}$, the predictive distribution for the objective vector is

$$\mathbf{y} \mid \mathcal{D},\, \mathbf{x} \;\sim\; \mathcal{N}\big(\boldsymbol{\mu},\, \boldsymbol{\Sigma}\big), \tag{4}$$

with mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$ and covariance $\boldsymbol{\Sigma}$. Throughout this thesis we restrict attention to the bi-objective case ($m = 2$), so that $\mathbf{y} \in \mathbb{R}^2$.

When each objective is modeled by an independent GP, the covariance matrix is diagonal: $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2,\, \sigma_2^2)$. When a multi-output GP is employed, for example via the convolution process framework [1, 7] or the linear model of coregionalization [2], the posterior admits a non-diagonal covariance

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\, \sigma_1 \sigma_2 \\ \rho\, \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}, \qquad |\rho| < 1, \tag{5}$$

where $\rho = \mathrm{Corr}(y_1, y_2)$ is the posterior correlation coefficient. The independent-objectives setting corresponds to $\rho = 0$.

2.2 Bivariate Gaussian Distribution

Definition 2.1 (Bivariate Gaussian)**.** A random vector $\mathbf{y} \in \mathbb{R}^2$ follows a non-degenerate bivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, written $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its probability density function is

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{2\pi\, |\boldsymbol{\Sigma}|^{1/2}} \exp\Big(-\tfrac{1}{2}\, (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\Big), \tag{6}$$

where $\boldsymbol{\Sigma}$ is symmetric positive definite, i.e. $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\top$ and $\mathbf{x}^\top \boldsymbol{\Sigma}\, \mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$ [20].

Proposition 2.2 (Explicit form for $m = 2$)**.** When $\boldsymbol{\Sigma}$ takes the form (5) with $|\rho| < 1$, we have

$$|\boldsymbol{\Sigma}| = \sigma_1^2\, \sigma_2^2\, (1 - \rho^2), \tag{7}$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_1^2\, \sigma_2^2\, (1 - \rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\, \sigma_1 \sigma_2 \\ -\rho\, \sigma_1 \sigma_2 & \sigma_1^2 \end{pmatrix}. \tag{8}$$

Proof. Directly computed using the formula for the inverse of a 2×2 matrix. □

Corollary 2.3 (Quadratic form)**.** Define the standardized quadratic form

$$Q(\mathbf{y}) := \frac{(y_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho\,(y_1-\mu_1)(y_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(y_2-\mu_2)^2}{\sigma_2^2}. \tag{9}$$

Then the exponent in the bivariate Gaussian PDF can be written as

$$-\tfrac{1}{2}\,(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}) = -\frac{Q(\mathbf{y})}{2(1-\rho^2)}. \tag{10}$$

Proof. Substitute $\boldsymbol{\Sigma}^{-1}$ from Proposition 2.2 and expand the quadratic form. □

Throughout the thesis we denote by $\varphi_{\mu,\sigma}(\cdot)$ and $\Phi_{\mu,\sigma}(\cdot)$ the PDF and CDF of a univariate Gaussian with mean μ and standard deviation σ, respectively. The standard normal PDF and CDF are written $\varphi(\cdot)$ and $\Phi(\cdot)$. For the bivariate case, $\varphi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\cdot)$ denotes the density (6), and $\Phi_2(a,b;\rho)$ denotes the standard bivariate normal CDF

$$\Phi_2(a,b;\rho) = \Pr(Z_1 \le a,\ Z_2 \le b), \qquad (Z_1,Z_2) \sim \mathcal{N}\Big(\mathbf{0},\ \left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)\Big). \tag{11}$$

2.3 Multivariate Gaussian Conditioning

The following standard result underpins both the treatment of correlated objectives in Chapter 3 and the analysis of correlated constraints in Chapter 4. It is stated here once for reference.

Lemma 2.4 (Affine transformation of Gaussian vectors)**.** Let $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and define $\mathbf{z} = A\mathbf{y} + \mathbf{b}$. Then $\mathbf{z} \sim \mathcal{N}(A\boldsymbol{\mu} + \mathbf{b},\ A\boldsymbol{\Sigma}A^\top)$.

Proof. Standard result from multivariate Gaussian theory [20]. □

Theorem 2.5 (Gaussian conditioning)**.** Suppose the random vector $(\mathbf{y}^\top, \mathbf{g}^\top)^\top$ is jointly Gaussian:

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{g}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_y\\ \boldsymbol{\mu}_g\end{pmatrix},\ \begin{pmatrix}\boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yg}\\ \boldsymbol{\Sigma}_{gy} & \boldsymbol{\Sigma}_{gg}\end{pmatrix}\right). \tag{12}$$

Then the conditional distribution of $\mathbf{y}$ given $\mathbf{g} = \mathbf{t}$ is

$$\mathbf{y} \mid \mathbf{g} = \mathbf{t} \ \sim\ \mathcal{N}\big(\boldsymbol{\mu}_{y|g}(\mathbf{t}),\ \boldsymbol{\Sigma}_{y|g}\big), \tag{13}$$

where

$$\boldsymbol{\mu}_{y|g}(\mathbf{t}) = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yg}\,\boldsymbol{\Sigma}_{gg}^{-1}\,(\mathbf{t}-\boldsymbol{\mu}_g), \tag{14}$$

$$\boldsymbol{\Sigma}_{y|g} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yg}\,\boldsymbol{\Sigma}_{gg}^{-1}\,\boldsymbol{\Sigma}_{gy}. \tag{15}$$

Note that $\boldsymbol{\Sigma}_{y|g}$ is independent of $\mathbf{t}$, while $\boldsymbol{\mu}_{y|g}(\mathbf{t})$ is affine in $\mathbf{t}$.

Proof. This is the standard Gaussian conditioning formula applied to the partitioning $(\mathbf{y}, \mathbf{g})$ of the joint distribution (12); see e.g. [20, Ch. 2.5.1]. □

Remark 2.6 (Induced dependence). Conditioning on a shared variable can introduce dependence between components that are marginally independent. For instance, if y_1 and y_2 are marginally independent ($\mathrm{Cov}(y_1, y_2) = 0$) but both correlate with a scalar g, then

$$\mathrm{Cov}(y_1, y_2 \mid g) = -\rho_1 \, \rho_2 \, \sigma_1 \, \sigma_2 \neq 0, \tag{16}$$

where $\rho_i = \mathrm{Corr}(y_i, g)$. This observation is central to the constrained analysis in Chapter 4.

Proof of (16). Under the assumption, the joint covariance of $(y_1, y_2, g)^\top$ is

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \rho_1\sigma_1\sigma_g \\ 0 & \sigma_2^2 & \rho_2\sigma_2\sigma_g \\ \rho_1\sigma_1\sigma_g & \rho_2\sigma_2\sigma_g & \sigma_g^2 \end{pmatrix}.$$

Partition as $(\mathbf{y}, g)$ with $\mathbf{y} = (y_1, y_2)^2\top$, giving

$$\Sigma_{\mathbf{yy}} = \mathrm{diag}(\sigma_1^2, \sigma_2^2), \quad \Sigma_{\mathbf{y}g} = (\rho_1\sigma_1\sigma_g,\ \rho_2\sigma_2\sigma_g)^\top, \quad \Sigma_{gg} = \sigma_g^2.$$

By Theorem 2.5,

$$\Sigma_{\mathbf{y}|g} = \Sigma_{\mathbf{yy}} - \Sigma_{\mathbf{y}g}\,\Sigma_{gg}^{-1}\,\Sigma_{g\mathbf{y}}.$$

The rank-one correction evaluates to

$$\Sigma_{\mathbf{y}g}\,\Sigma_{gg}^{-1}\,\Sigma_{g\mathbf{y}} = \begin{pmatrix} \rho_1^2\sigma_1^2 & \rho_1\rho_2\sigma_1\sigma_2 \\ \rho_1\rho_2\sigma_1\sigma_2 & \rho_2^2\sigma_2^2 \end{pmatrix},$$

such that

$$\Sigma_{\mathbf{y}|g} = \begin{pmatrix} \sigma_1^2(1-\rho_1^2) & -\rho_1\rho_2\sigma_1\sigma_2 \\ -\rho_1\rho_2\sigma_1\sigma_2 & \sigma_2^2(1-\rho_2^2) \end{pmatrix}.$$

The off-diagonal entry is $\mathrm{Cov}(y_1, y_2 \mid g) = -\rho_1\rho_2\sigma_1\sigma_2$ which is non-zero whenever $\rho_1 \neq 0$ and $\rho_2 \neq 0$. □

2.4 Hypervolume and Generalized Hypervolume Improvement

The following introduces the hypervolume indicator and three variants of the hypervolume improvement that are used as building blocks for acquisition functions in multi-objective Bayesian optimization.

2.4.1 Pareto Dominance and the Hypervolume Indicator

For every $\mathbf{y}^{(1)}, \mathbf{y}^{(2)} \in \mathbb{R}^m$, we say $\mathbf{y}^{(1)}$ *weakly dominates* $\mathbf{y}^{(2)}$ (written $\mathbf{y}^{(1)} \preceq \mathbf{y}^{(2)}$) if and only if $y_i^{(1)} \leq y_i^{(2)}$ for all $i \in [1..m]$. The *Pareto order* $\prec$ is defined by $\mathbf{y}^{(1)} \prec \mathbf{y}^{(2)}$ if and only if $\mathbf{y}^{(1)} \preceq \mathbf{y}^{(2)}$ and $\mathbf{y}^{(1)} \neq \mathbf{y}^{(2)}$. A point $\mathbf{x} \in \mathcal{X}$ is *efficient* if and only if there is no $\mathbf{x}' \in \mathcal{X}$ with $\mathbf{f}(\mathbf{x}') \prec \mathbf{f}(\mathbf{x})$. The image of the efficient set under $\mathbf{f}$ is the *Pareto front*.

Multi-objective optimization algorithms maintain a finite approximation set $\mathcal{P}$, which is the non-dominated subset of the evaluated objective vectors $\mathbf{Y}$. The *non-dominated space* with respect to $\mathcal{P}$ is $\mathrm{ndom}(\mathcal{P}) := \{\mathbf{y} \in \mathbb{R}^m : \nexists\, \mathbf{p} \in \mathcal{P}\ (\mathbf{p} \prec \mathbf{y})\}$, and the *dominated space* is $\mathrm{dom}(\mathcal{P}) := \mathbb{R}^m \setminus \mathrm{ndom}(\mathcal{P})$.

Definition 2.7 (Hypervolume indicator [32])**.** For a finite set $\mathcal{P} \subset \mathbb{R}^m$ and a reference point $\mathbf{r} \in \mathbb{R}^m$, the hypervolume indicator is

$$\mathrm{HV}(\mathcal{P}, \mathbf{r}) = \lambda\big(\{\mathbf{y} \in \mathbb{R}^m : \exists\, \mathbf{p} \in \mathcal{P},\ \mathbf{p} \prec \mathbf{y} \prec \mathbf{r}\}\big), \tag{17}$$

where λ denotes the Lebesgue measure.

2.4.2 Positive, Negative, and Generalized Hypervolume Improvement

Definition 2.8 (Hypervolume improvement [12, 27])**.** For a point $\mathbf{y} \in \mathbb{R}^2$, the *(positive) hypervolume improvement* with respect to $\mathcal{P}$ and $\mathbf{r}$ is

$$\Delta^+(\mathbf{y}; \mathcal{P}, \mathbf{r}) = \mathrm{HV}(\mathcal{P} \cup \{\mathbf{y}\}, \mathbf{r}) - \mathrm{HV}(\mathcal{P}, \mathbf{r}). \tag{18}$$

Note that $\Delta^+(\mathbf{y}) \geq 0$ for all $\mathbf{y}$, with equality if and only if $\mathbf{y}$ is dominated by a point in $\mathcal{P}$ or $\mathbf{y} \not\prec \mathbf{r}$.

Definition 2.9 (Negative hypervolume improvement [27])**.** For a dominated point $\mathbf{y} \in \mathrm{dom}(\mathcal{P})$, the *negative hypervolume "improvement"* is

$$\Delta^-(\mathbf{y}; \mathcal{P}) = -\lambda\big(\{\mathbf{p} \in \mathbb{R}^2 : \mathbf{p} \preceq \mathbf{y}\} \cap \mathrm{dom}(\mathcal{P})\big). \tag{19}$$

The negative HVI penalises dominated points proportionally to how deeply they lie inside the dominated region. Points farther from the Pareto front receive a more negative value.

Definition 2.10 (Generalized hypervolume improvement [27])**.** The *generalized hypervolume improvement* combines the positive and negative cases:

$$\Delta(\mathbf{y}) = \mathbf{1}_{\mathrm{ndom}(\mathcal{P})}(\mathbf{y})\, \Delta^+(\mathbf{y}) + \mathbf{1}_{\mathrm{dom}(\mathcal{P})}(\mathbf{y})\, \Delta^-(\mathbf{y}). \tag{20}$$

This definition assigns a continuous real value to every point in the objective space: positive for non-dominated points (measuring the area gained), negative for dominated points (measuring the depth of domination), and zero exactly on the attainment boundary.

2.5 Cell Partition Framework

Following [27], the objective space is partitioned into axis-aligned cells determined by the approximation set $\mathcal{P}$. Within each cell, the hypervolume improvement admits a closed-form expression as a product of translated coordinates plus a cell-specific constant. All geometric constructions in this subsection are for the bi-objective case ($m = 2$).

2.5.1 Augmented Pareto Set and Cell Decomposition

Definition 2.11 (Augmented Pareto set [27])**.** Let $\mathcal{P} = \{\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(n)}\}$ be a set of mutually non-dominated points with $\mathbf{p}^{(k)} \prec \mathbf{r}$ for every k, and assume the first coordinates $p_1^{(1)}, \ldots, p_1^{(n)}$ are distinct. Define the augmented set

$$\widetilde{\mathcal{P}} = \mathcal{P} \;\cup\; \big\{(-\infty,\, r_2)^\top,\; (r_1,\, -\infty)^\top\big\}. \tag{21}$$

Index the $n+2$ points of $\widetilde{\mathcal{P}}$ as $\mathbf{y}^{(0)}, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n+1)}$ in increasing order of their first coordinate:

$$y_1^{(0)} < y_1^{(1)} < \cdots < y_1^{(n+1)},$$

where $y_1^{(0)} = -\infty$ and $y_1^{(n+1)} = r_1$. Because the points of $\mathcal{P}$ are mutually non-dominated, the second coordinates are strictly decreasing:

$$y_2^{(0)} > y_2^{(1)} > \cdots > y_2^{(n+1)},$$

where $y_2^{(0)} = r_2$ and $y_2^{(n+1)} = -\infty$. The augmented set thus forms a descending staircase from $(-\infty, r_2)$ to $(r_1, -\infty)$ in the objective space.

Definition 2.12 (Cell partition [27])**.** The coordinates of $\widetilde{\mathcal{P}}$ induce a grid of axis-aligned rectangles covering $(-\infty, r_1] \times (-\infty, r_2]$. For indices $i, j \in \{0, 1, \ldots, n\}$, the cell $C(i, j)$ is the rectangle

$$C(i,j) = \left[y_1^{(i)},\ y_1^{(i+1)}\right] \times \left[y_2^{(n-j+1)},\ y_2^{(n-j)}\right]. \tag{22}$$

Here i indexes the vertical strips from left to right, and j indexes the horizontal strips from bottom to top.

Proposition 2.13. The $(n+1)^2$ cells $\{C(i,j)\}_{i,j=0}^{n}$ partition the region $(-\infty, r_1] \times (-\infty, r_2]$, and distinct cells are disjoint up to a set of Lebesgue measure zero.

Proof. The first coordinates $y_1^{(0)} < \cdots < y_1^{(n+1)}$ subdivide $(-\infty, r_1]$ into $n+1$ consecutive intervals $[y_1^{(i)}, y_1^{(i+1)}]$, $i = 0, \ldots, n$. Likewise, the second coordinates $y_2^{(n+1)} < \cdots < y_2^{(0)}$ subdivide $(-\infty, r_2]$ into $n+1$ consecutive intervals $[y_2^{(n-j+1)}, y_2^{(n-j)}]$, $j = 0, \ldots, n$. The cells are exactly the products of one interval from each family, so they tile the full rectangle. Adjacent cells share at most an edge or a vertex, so their interiors are disjoint. □

Definition 2.14 (Cell corners)**.** For each cell $C(i, j)$, define its lower-left and upper-right corners:

$$\mathbf{l}^{(i,j)} = \begin{pmatrix} y_1^{(i)} \\ y_2^{(n-j+1)} \end{pmatrix}, \qquad \mathbf{u}^{(i,j)} = \begin{pmatrix} y_1^{(i+1)} \\ y_2^{(n-j)} \end{pmatrix}. \tag{23}$$

Under the minimization convention, "lower-left" corresponds to the better (smaller) objective values and "upper-right" to the worse (larger) values.

Lemma 2.15 (Non-dominated vs. dominated cells)**.** A cell $C(i, j)$ lies in the non-dominated region with respect to $\mathcal{P}$ if and only if $i + j \le n$. Equivalently, the cell is dominated by some point in $\mathcal{P}$ if and only if $i + j \ge n + 1$.

Proof. A point $\mathbf{y}$ in the interior of $C(i, j)$ satisfies $y_1^{(i)} < y_1 < y_1^{(i+1)}$ and $y_2^{(n-j+1)} < y_2 < y_2^{(n-j)}$. The point $\mathbf{y}$ is dominated by some $\mathbf{p}^{(k)} \in \mathcal{P}$ if and only if there exists $k \in \{1, \ldots, n\}$ with (1) $k \le i$ and (2) $k \ge n - j + 1$. Such k exists if and only if $n - j + 1 \le i$, i.e. $i + j \ge n + 1$. □

2.5.2 Cell-wise HVI Structure

Theorem 2.16 (Cell-wise HVI: non-dominated cells)**.** Let $C(i, j)$ be a non-dominated cell (i.e. $i + j \le n$). For any $\mathbf{y} \in C(i, j)$, the hypervolume improvement can be expressed as

$$\Delta^+(\mathbf{y})\big|_{C(i,j)} = z_1\, z_2 + \gamma^{(i,j)}, \tag{24}$$

where

$$z_1 = y_1^{(n-j+1)} - y_1 \ge 0, \qquad z_2 = y_2^{(i)} - y_2 \ge 0, \tag{25}$$

are the signed distances from $\mathbf{y}$ to the attainment boundary of $\mathcal{P}$ along each coordinate axis, and

$$\gamma^{(i,j)} = \Delta^+(\mathbf{u}^{(i,j)}) - \left(y_1^{(n-j+1)} - u_1^{(i,j)}\right)\left(y_2^{(i)} - u_2^{(i,j)}\right) \tag{26}$$

is a constant depending only on the cell, not on $\mathbf{y}$.

Proof. We construct the hypervolume improvement region explicitly and decompose it into four rectangles.

Step 1: The HVI region. The hypervolume improvement equals the Lebesgue measure of $S(\mathbf{y}) = \mathrm{dom}(\{\mathbf{y}\}) \setminus \mathrm{dom}(\mathcal{P})$.

Step 2: Decompose into four disjoint regions. Write $\mathbf{u} = \mathbf{u}^{(i,j)} = (u_1, u_2)$:

$$A_1 = [y_1,\, u_1] \times [y_2,\, u_2], \tag{27}$$

$$A_2 = [u_1,\, y_1^{(n-j+1)}] \times [y_2,\, u_2], \tag{28}$$

$$A_3 = [y_1,\, u_1] \times [u_2,\, y_2^{(i)}], \tag{29}$$

$$A_4 = \mathrm{dom}\big(\{\mathbf{u}\}\big) \setminus \mathrm{dom}(\mathcal{P}). \tag{30}$$

Step 3: Compute the areas.

$$\lambda(A_1) = (u_1 - y_1)(u_2 - y_2), \qquad \lambda(A_2) = (y_1^{(n-j+1)} - u_1)(u_2 - y_2), \tag{31}$$

$$\lambda(A_3) = (u_1 - y_1)(y_2^{(i)} - u_2), \qquad \lambda(A_4) = \Delta^+(\mathbf{u}^{(i,j)}). \tag{32}$$

Step 4: Sum and simplify. Adding $\lambda(A_1)$ and $\lambda(A_2)$:

$$\lambda(A_1) + \lambda(A_2) = (y_1^{(n-j+1)} - y_1)(u_2 - y_2). \tag{33}$$

Expanding and collecting terms:

$$\Delta^+(\mathbf{y}) = \underbrace{(y_1^{(n-j+1)} - y_1)(y_2^{(i)} - y_2)}_{=z_1\, z_2} + \underbrace{\Delta^+(\mathbf{u}) - (y_1^{(n-j+1)} - u_1)(y_2^{(i)} - u_2)}_{=\gamma^{(i,j)}}. \tag{34}$$

□

Remark 2.17 (Staircase projections). The attainment boundary of $\mathcal{P}$ is a descending staircase passing through $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}$. For a point $\mathbf{y}$ in cell $C(i,j)$: projecting $\mathbf{y}$ horizontally onto the staircase hits at $y_1 = y_1^{(n-j+1)}$, and projecting $\mathbf{y}$ vertically onto the staircase hits at $y_2 = y_2^{(i)}$. Since $\mathbf{y}$ lies in the non-dominated region, both projections point towards larger coordinate values, giving $z_1, z_2 \geq 0$. The product $z_1 z_2$ measures the area of the rectangle between $\mathbf{y}$ and its staircase projections, and vanishes when $\mathbf{y}$ lies on the staircase itself.

Corollary 2.18 (Range of HVI within a non-dominated cell)**.** For a non-dominated cell $C(i,j)$, $\Delta^+(\mathbf{y}) \in [\Delta^+(\mathbf{u}^{(i,j)}),\ \Delta^+(\mathbf{l}^{(i,j)})]$ for all $\mathbf{y} \in C(i,j)$.

Proof. Within the cell, z_1 and z_2 are non-negative and increase as $\mathbf{y}$ moves from $\mathbf{u}^{(i,j)}$ towards $\mathbf{l}^{(i,j)}$. Since $\Delta^+ = z_1 z_2 + \gamma^{(i,j)}$ and $\gamma^{(i,j)}$ is constant, the extrema are at the corners. □

Theorem 2.19 (Cell-wise HVI: dominated cells)**.** Let $C(i,j)$ be a dominated cell (i.e. $i + j \geq n+1$). For any $\mathbf{y} \in C(i,j)$, the negative hypervolume improvement can be expressed as

$$\Delta^-(\mathbf{y})\big|_{C(i,j)} = -\bar{z}_1\, \bar{z}_2 + \gamma^{-(i,j)}, \tag{35}$$

where

$$\bar{z}_1 = y_1 - y_1^{(n-j+1)} \geq 0, \qquad \bar{z}_2 = y_2 - y_2^{(i)} \geq 0, \tag{36}$$

are the signed distances from $\mathbf{y}$ to the attainment boundary measured into the dominated region, and

$$\gamma^{-(i,j)} = \Delta^-(\mathbf{l}^{(i,j)}) + \big(l_1^{(i,j)} - y_1^{(n-j+1)}\big)\big(l_2^{(i,j)} - y_2^{(i)}\big) \tag{37}$$

is a constant depending only on the cell.

Proof. The proof follows by an analogous geometric decomposition to Theorem 2.16, applied to the dominated region.

Step 1. By Definition 2.9, $|\Delta^-(\mathbf{y})|$ equals the Lebesgue measure of $S^-(\mathbf{y}) = \{\mathbf{p} \in \mathbb{R}^2 : \mathbf{p} \preceq \mathbf{y}\} \cap \mathrm{dom}(\mathcal{P})$.

Step 2. Write $\mathbf{l} = \mathbf{l}^{(i,j)} = (l_1, l_2)$. Since $i + j \geq n + 1$, the entire cell lies in the dominated region. We split $S^-(\mathbf{y})$:

$$A'_1 = [l_1,\, y_1] \times [l_2,\, y_2], \tag{38}$$
$$A'_2 = [y_1^{(n-j+1)},\, l_1] \times [l_2,\, y_2], \tag{39}$$
$$A'_3 = [l_1,\, y_1] \times [y_2^{(i)},\, l_2], \tag{40}$$
$$A'_4 = \{\mathbf{p} \in \mathbb{R}^2 : \mathbf{p} \preceq \mathbf{l}\} \cap \mathrm{dom}(\mathcal{P}). \tag{41}$$

Step 3.

$$\lambda(A'_1) = (y_1 - l_1)(y_2 - l_2), \qquad \lambda(A'_2) = (l_1 - y_1^{(n-j+1)})(y_2 - l_2), \tag{42}$$
$$\lambda(A'_3) = (y_1 - l_1)(l_2 - y_2^{(i)}), \qquad \lambda(A'_4) = |\Delta^-(\mathbf{l}^{(i,j)})|. \tag{43}$$

Step 4. Adding and simplifying:

$$|\Delta^-(\mathbf{y})| = \underbrace{(y_1 - y_1^{(n-j+1)})(y_2 - y_2^{(i)})}_{=\bar{z}_1\,\bar{z}_2} + \underbrace{|\Delta^-(\mathbf{l})| - (l_1 - y_1^{(n-j+1)})(l_2 - y_2^{(i)})}_{=|\gamma^{-(i,j)}|}. \tag{44}$$

Since $\Delta^-(\mathbf{y}) = -|\Delta^-(\mathbf{y})|$ by definition, $\Delta^-(\mathbf{y})\big|_{C(i,j)} = -\bar{z}_1\,\bar{z}_2 + \gamma^{-(i,j)}$ as claimed. □

Remark 2.20 (Coordinate inversion for dominated cells). Following [27], the dominated-cell formula can be related to the non-dominated case by a coordinate inversion. Note that $\bar{z}_k = -z_k$ for $k = 1, 2$, where z_k are defined as in (25). In a dominated cell $\bar{z}_1, \bar{z}_2 \geq 0$. The negative HVI takes the form $\Delta^- = -\bar{z}_1\bar{z}_2 + \gamma^{-(i,j)}$, structurally identical to $\Delta^+ = z_1 z_2 + \gamma^{(i,j)}$ up to a sign flip on the product term. This observation is exploited when deriving the probability distribution.

Corollary 2.21 (Range of HVI within a dominated cell)**.** For a dominated cell $C(i, j)$ with $i + j \geq n + 1$, $\Delta^-(\mathbf{y}) \in [\Delta^-(\mathbf{u}^{(i,j)}),\ \Delta^-(\mathbf{l}^{(i,j)})]$ for all $\mathbf{y} \in C(i, j)$, where $\Delta^-(\mathbf{u}^{(i,j)}) \leq \Delta^-(\mathbf{l}^{(i,j)}) \leq 0$.

Proof. $\Delta^- = -\bar{z}_1\bar{z}_2 + \gamma^{-(i,j)}$ is minimized when $\bar{z}_1\bar{z}_2$ is largest (at $\mathbf{u}^{(i,j)}$) and maximized when smallest (at $\mathbf{l}^{(i,j)}$). □

Corollary 2.22 (Unified generalized HVI)**.** For any cell $C(i, j)$, the generalized hypervolume improvement (Definition 2.10) can be written as

$$\Delta(\mathbf{y})\big|_{C(i,j)} = \begin{cases} z_1\, z_2 + \gamma^{(i,j)}, & i + j \leq n, \\ -\bar{z}_1\, \bar{z}_2 + \gamma^{-(i,j)}, & i + j \geq n + 1. \end{cases} \tag{45}$$

2.5.3 Cell Probabilities

When the objective vector is a Gaussian random variable $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we require the probability that $\mathbf{y}$ falls in each cell and the distribution of $\Delta(\mathbf{y})$ within that cell.

Definition 2.23 (Truncation bounds: non-dominated cells)**.** For a non-dominated cell $C(i,j)$ $(i+j \le n)$, define

$$L_1 = y_1^{(n-j+1)} - u_1^{(i,j)}, \qquad U_1 = y_1^{(n-j+1)} - l_1^{(i,j)}, \tag{46}$$

$$L_2 = y_2^{(i)} - u_2^{(i,j)}, \qquad U_2 = y_2^{(i)} - l_2^{(i,j)}. \tag{47}$$

The truncation region is $\mathcal{R}^{(i,j)} = [L_1, U_1] \times [L_2, U_2]$.

Definition 2.24 (Truncation bounds: dominated cells)**.** For a dominated cell $C(i,j)$ $(i+j \ge n+1)$, define

$$\bar{L}_1 = l_1^{(i,j)} - y_1^{(n-j+1)}, \qquad \bar{U}_1 = u_1^{(i,j)} - y_1^{(n-j+1)}, \tag{48}$$

$$\bar{L}_2 = l_2^{(i,j)} - y_2^{(i)}, \qquad \bar{U}_2 = u_2^{(i,j)} - y_2^{(i)}. \tag{49}$$

The truncation region is $\bar{\mathcal{R}}^{(i,j)} = [\bar{L}_1, \bar{U}_1] \times [\bar{L}_2, \bar{U}_2]$.

Proposition 2.25 (Non-negativity of truncation bounds)**.** For a non-dominated cell $(i+j \le n)$: $0 \le L_1 \le U_1$ and $0 \le L_2 \le U_2$. For a dominated cell $(i+j \ge n+1)$: $0 \le \bar{L}_1 \le \bar{U}_1$ and $0 \le \bar{L}_2 \le \bar{U}_2$.

Proof. Non-dominated case. $L_1 = y_1^{(n-j+1)} - y_1^{(i+1)}$ and $U_1 = y_1^{(n-j+1)} - y_1^{(i)}$. Since $i+j \le n$ implies $i+1 \le n-j+1$, the strict ordering gives $0 \le L_1 \le U_1$. The argument for L_2, U_2 is analogous.

Dominated case. $\bar{L}_1 = y_1^{(i)} - y_1^{(n-j+1)}$ and $\bar{U}_1 = y_1^{(i+1)} - y_1^{(n-j+1)}$. Since $i+j \ge n+1$ implies $i \ge n-j+1$, we have $y_1^{(i)} \ge y_1^{(n-j+1)}$, giving $0 \le \bar{L}_1 \le \bar{U}_1$. Note that $\bar{L}_1 = 0$ if and only if $i+j = n+1$ (diagonal cells adjacent to the staircase). The argument for $\bar{L}_2, \bar{U}_2$ is analogous. □

Proposition 2.26 (Distribution of transformed coordinates: non-dominated cells)**.** If $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the transformation $z_1 = y_1^{(n-j+1)} - y_1$, $z_2 = y_2^{(i)} - y_2$ yields

$$\mathbf{z} = (z_1, z_2)^\top \sim \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}), \qquad \boldsymbol{\mu}' = \begin{pmatrix} y_1^{(n-j+1)} - \mu_1 \\ y_2^{(i)} - \mu_2 \end{pmatrix}. \tag{50}$$

In particular, the covariance matrix is unchanged.

Proof. Write $\mathbf{z} = (-I)\mathbf{y} + (y_1^{(n-j+1)}, y_2^{(i)})^\top$. By Lemma 2.4, $A\boldsymbol{\Sigma}A^\top = (-I)\boldsymbol{\Sigma}(-I)^\top = \boldsymbol{\Sigma}$ and the mean is $-\boldsymbol{\mu} + \mathbf{b} = \boldsymbol{\mu}'$. □

Proposition 2.27 (Distribution of transformed coordinates: dominated cells)**.** For a dominated cell $C(i,j)$ $(i+j \ge n+1)$, the transformation $\bar{z}_1 = y_1 - y_1^{(n-j+1)}$, $\bar{z}_2 = y_2 - y_2^{(i)}$ yields

$$\bar{\mathbf{z}} = (\bar{z}_1, \bar{z}_2)^\top \sim \mathcal{N}(\boldsymbol{\mu}^-, \boldsymbol{\Sigma}), \qquad \boldsymbol{\mu}^- = \begin{pmatrix} \mu_1 - y_1^{(n-j+1)} \\ \mu_2 - y_2^{(i)} \end{pmatrix} = -\boldsymbol{\mu}'. \tag{51}$$

Again, the covariance matrix is unchanged.

Proof. Write $\bar{\mathbf{z}} = I\,\mathbf{y} - (y_1^{(n-j+1)},\, y_2^{(i)})^\top$. By Lemma 2.4, $I\boldsymbol{\Sigma}I^\top = \boldsymbol{\Sigma}$ and the mean is $\boldsymbol{\mu} + \mathbf{b}^- = \boldsymbol{\mu}^-$. □

Remark 2.28 (Covariance preservation). Both transformations preserve $\boldsymbol{\Sigma}$. In particular, if $\rho \neq 0$, then z_1, z_2 (or $\bar{z}_1, \bar{z}_2$) remain correlated with the same correlation coefficient. This is the key difference from the independent-objectives setting of [27]: when $\rho \neq 0$, $\mathbb{E}[z_1 z_2] \neq \mathbb{E}[z_1]\,\mathbb{E}[z_2]$.

Definition 2.29 (Cell probability)**.**

$$P_{\text{cell}}^{(i,j)} = \Pr(\mathbf{y} \in C(i,j)) = \int_{\mathcal{R}^{(i,j)}} \varphi_{\boldsymbol{\mu}',\boldsymbol{\Sigma}}(z_1, z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2 \tag{52}$$

for non-dominated cells (equivalently, $\int_{\bar{\mathcal{R}}^{(i,j)}} \varphi_{\boldsymbol{\mu}^-,\boldsymbol{\Sigma}}(\bar{z}_1, \bar{z}_2)\,\mathrm{d}\bar{z}_1\,\mathrm{d}\bar{z}_2$ for dominated cells).

When $\rho = 0$, the cell probability factors as a product of univariate CDF differences [27]. The closed-form expression for general $\rho \neq 0$ is derived in Chapter 3 (Theorem 3.6).

2.6 Distribution of a Product of Two Variables

By the cell-wise structure established above, the distribution of generalized HVI within each cell is determined by the distribution of a product of (possibly sign-flipped) truncated correlated Gaussians. The following lemma and theorems provide the required density formulas.

Lemma 2.30 (PDF of a product via change of variables)**.** Let (X, Y) have joint density $f_{X,Y}(x, y)$. Then $Z = XY$ has density

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}\Big(x,\, \frac{z}{x}\Big)\,\frac{1}{|x|}\,\mathrm{d}x. \tag{53}$$

Proof. Apply the transformation $(x, y) \mapsto (u, v) = (x,\, xy)$ with Jacobian $|\det J| = |x|$ and marginalise over u. □

Theorem 2.31 (Product density: non-dominated cells)**.** Let (z_1, z_2) follow the truncated bivariate normal on $\mathcal{R}^{(i,j)} = [L_1, U_1] \times [L_2, U_2]$ with $0 \leq L_k \leq U_k$. Define $Z = z_1 z_2$. Then:

$$f_Z^{\text{trunc}}(z) = \frac{1}{P_{\text{cell}}^{(i,j)}} \int_{\alpha(z)}^{\beta(z)} \varphi_{\boldsymbol{\mu}',\boldsymbol{\Sigma}}\Big(x,\, \frac{z}{x}\Big)\,\frac{1}{x}\,\mathrm{d}x, \qquad z \in [L_1 L_2,\, U_1 U_2], \tag{54}$$

and $f_Z^{\text{trunc}}(z) = 0$ otherwise, with integration bounds

$$\alpha(z) = \max\Big(L_1,\, \frac{z}{U_2}\Big), \qquad \beta(z) = \begin{cases} \min\big(U_1,\, \frac{z}{L_2}\big), & L_2 > 0, \\ U_1, & L_2 = 0. \end{cases} \tag{55}$$

Proof. Since $L_1, L_2 \geq 0$, we have $z_1, z_2 \geq 0$ a.s. under truncation, so $Z \geq 0$ and $|x| = x$. Applying Lemma 2.30 to the truncated density and enforcing $x \in [L_1, U_1]$, $z/x \in [L_2, U_2]$ yields the bounds. □

Theorem 2.32 (Product density: dominated cells)**.** Let $(\bar{z}_1, \bar{z}_2)$ follow the truncated bivariate normal on $\bar{\mathcal{R}}^{(i,j)} = [\bar{L}_1, \bar{U}_1] \times [\bar{L}_2, \bar{U}_2]$ with $0 \le \bar{L}_k \le \bar{U}_k$. Define $\bar{Z} = \bar{z}_1 \bar{z}_2$. Then:

$$f_{\bar{Z}}^{\text{trunc}}(\bar{z}) = \frac{1}{P_{\text{cell}}^{(i,j)}} \int_{\bar{\alpha}(\bar{z})}^{\bar{\beta}(\bar{z})} \varphi_{\boldsymbol{\mu}^-, \boldsymbol{\Sigma}}\Big(x, \frac{\bar{z}}{x}\Big) \frac{1}{x} \, \mathrm{d}x, \qquad \bar{z} \in [\bar{L}_1 \bar{L}_2, \ \bar{U}_1 \bar{U}_2], \tag{56}$$

and zero otherwise, with

$$\bar{\alpha}(\bar{z}) = \max\Big(\bar{L}_1, \frac{\bar{z}}{\bar{U}_2}\Big), \qquad \bar{\beta}(\bar{z}) = \begin{cases} \min\big(\bar{U}_1, \frac{\bar{z}}{\bar{L}_2}\big), & \bar{L}_2 > 0, \\ \bar{U}_1, & \bar{L}_2 = 0. \end{cases} \tag{57}$$

Proof. Identical to Theorem 2.31 with $(\boldsymbol{\mu}', L_k, U_k)$ replaced by $(\boldsymbol{\mu}^-, \bar{L}_k, \bar{U}_k)$. □

Proposition 2.33 (Laurent polynomial structure of the exponent)**.** The quadratic form evaluated at $(x, \, z/x)$ takes the form

$$Q\Big(x, \frac{z}{x}\Big) = a_2 \, x^2 + a_1 \, x + a_0 + b_1 \, x^{-1} + b_2 \, x^{-2}, \tag{58}$$

where the coefficients (using any shifted mean $\boldsymbol{\nu}$, i.e. $\boldsymbol{\mu}'$ for non-dominated or $\boldsymbol{\mu}^-$ for dominated cells) are:

$$a_2 = \frac{1}{\sigma_1^2}, \tag{59}$$

$$a_1 = -\frac{2\nu_1}{\sigma_1^2} + \frac{2\rho \, \nu_2}{\sigma_1 \sigma_2}, \tag{60}$$

$$a_0 = \frac{\nu_1^2}{\sigma_1^2} + \frac{\nu_2^2}{\sigma_2^2} - \frac{2\rho \, z}{\sigma_1 \sigma_2} - \frac{2\rho \, \nu_1 \nu_2}{\sigma_1 \sigma_2}, \tag{61}$$

$$b_1 = -\frac{2\nu_2 \, z}{\sigma_2^2} + \frac{2\rho \, \nu_1 \, z}{\sigma_1 \sigma_2}, \tag{62}$$

$$b_2 = \frac{z^2}{\sigma_2^2}. \tag{63}$$

Proof. Substitute $z_1 = x$, $z_2 = z/x$ into the quadratic form of (9) (with mean $\boldsymbol{\nu}$) and expand each term. The cross term $-\frac{2\rho}{\sigma_1 \sigma_2}(x - \nu_1)(z/x - \nu_2)$ generates terms coupling x and z. □

Remark 2.34 (Effect of correlation). Setting $\rho = 0$ eliminates the coupling terms: $a_1 \to -2\nu_1/\sigma_1^2$, the ρ-dependent parts of a_0 and b_1 vanish, and the exponent separates into a function of x alone plus a function of z/x alone. When $\rho \neq 0$, this separation fails, requiring genuinely two-dimensional integration.

Remark 2.35 (Identical structure for dominated cells). Since the only difference between the non-dominated and dominated product densities is the shifted mean ($\boldsymbol{\mu}'$ vs. $\boldsymbol{\mu}^-$), the Laurent polynomial structure (58) applies to both cases with the appropriate substitution of $\boldsymbol{\nu}$.

3 Extension I: Correlated Gaussian Objectives

This chapter develops the first main contribution of the thesis: the exact probability distribution of the generalized hypervolume improvement under correlated bivariate Gaussian posteriors. The presentation builds exclusively on the definitions and tools established in Chapter 2; no constraints are introduced here. Those are deferred to Chapter 4.

3.1 From Independent to Correlated Objectives

Wang et al. [27] derive the distribution of the generalized hypervolume improvement (HVI) random variable $\Delta(\boldsymbol{y})$ under the assumption that the two objective posteriors are *independent*:

$$y_1 \mid \mathcal{D}, \boldsymbol{x} \sim \mathcal{N}(\mu_1, \sigma_1^2), \qquad y_2 \mid \mathcal{D}, \boldsymbol{x} \sim \mathcal{N}(\mu_2, \sigma_2^2), \qquad y_1 \perp y_2 \mid \mathcal{D}. \tag{64}$$

Under this assumption, the joint density of (y_1, y_2) factors as a product of two univariate Gaussian densities. This factorization propagates through the entire derivation: cell probabilities decompose into products of univariate CDF differences, and the product density used to compute the within-cell HVI distribution separates into a one-dimensional integral involving the product of two independent univariate normal densities [27, Eq. 4].

In practice, however, objective independence is often violated. Multi-output Gaussian process models, such as the linear model of coregionalization [2], convolution processes [7], or intrinsic coregionalization models [6], induce a joint posterior in which the predictive distributions of $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ are correlated. Even when separate GPs are fitted to each objective, shared input features or overlapping training data can produce non-negligible posterior correlations via multi-task learning schemes. The correlated bivariate Gaussian posterior takes the form

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \;\Big|\; \mathcal{D}, \boldsymbol{x} \;\sim\; \mathcal{N}\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}\Big), \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \tag{65}$$

where $\rho = \mathrm{Corr}(y_1, y_2) \in (-1, 1)$ is the posterior correlation coefficient. When $\rho = 0$, the covariance matrix $\boldsymbol{\Sigma}$ is diagonal and (65) reduces to the independent setting (64).

The goal of this chapter is to derive the cumulative distribution function and probability density function of $\Delta(\boldsymbol{y})$ under the general posterior (65), using the cell-partition framework and the generalized HVI definition recalled in Sections 2.4.2 and 2.5 of Chapter 2. The key analytical challenge is that, once $\rho \neq 0$, the product structure exploited by Wang et al. no longer holds: cell probabilities become bivariate normal rectangle probabilities, and the within-cell product density involves the full bivariate Gaussian density rather than a product of marginals. In Section 3.3 it is shown that the cell probabilities require evaluation of the bivariate normal CDF, and in Section 3.4 the product density is derived under correlation and identify its characteristic Laurent polynomial exponent structure. Section 3.5 assembles the complete marginal distribution as a mixture over cells and proves that it reduces exactly to the results of Wang et al. when $\rho = 0$.

3.2 Distribution of Transformed Coordinates

The cell-partition framework (Section 2.5) expresses the HVI within each cell in terms of transformed coordinates. For a non-dominated cell $C(i, j)$ with $i + j \leq n$, one defines

$$z_1 = y_1^{(n-j+1)} - y_1, \qquad z_2 = y_2^{(i)} - y_2, \tag{66}$$

so that the positive HVI restricted to the cell is $\Delta^+(\boldsymbol{y})\big|_{C(i,j)} = z_1 z_2 + \gamma^{(i,j)}$ (Section 2.4.2). For a dominated cell $C(i,j)$ with $i+j \geq n+1$, the analogous transformation is

$$\bar{z}_1 = y_1 - y_1^{(n-j+1)}, \qquad \bar{z}_2 = y_2 - y_2^{(i)}, \tag{67}$$

yielding $\Delta^-(\boldsymbol{y})\big|_{C(i,j)} = -\bar{z}_1 \bar{z}_2 + \gamma^{-(i,j)}$.

In both cases, the transformation is affine in $\boldsymbol{y}$. The following results establish that the covariance structure is preserved, a property that is immediate under independence but that must be stated explicitly in the correlated setting, since it is the base of every subsequent derivation.

Lemma 3.1 (Affine transformation of Gaussian vectors)**.** Let $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and define $\boldsymbol{z} = A\boldsymbol{y} + \boldsymbol{b}$ for a constant matrix $A \in \mathbb{R}^{2\times 2}$ and vector $\boldsymbol{b} \in \mathbb{R}^2$. Then $\boldsymbol{z} \sim \mathcal{N}(A\boldsymbol{\mu} + \boldsymbol{b},\ A\boldsymbol{\Sigma}A^\top)$.

Proof. Standard result from multivariate Gaussian theory; see, e.g., Koch [20]. □

This lemma is now applied to each type of cell.

Proposition 3.2 (Transformed coordinates: non-dominated cells)**.** Let $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ as in (65). The transformation (66) yields

$$\boldsymbol{z} = (z_1, z_2)^\top \sim \mathcal{N}\big(\boldsymbol{\mu}',\ \boldsymbol{\Sigma}\big), \qquad \boldsymbol{\mu}' = \begin{pmatrix} y_1^{(n-j+1)} - \mu_1 \\ y_2^{(i)} - \mu_2 \end{pmatrix}. \tag{68}$$

In particular, the covariance matrix is unchanged: $\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{\Sigma}$.

Proof. Write $\boldsymbol{z} = (-I)\,\boldsymbol{y} + \big(y_1^{(n-j+1)},\, y_2^{(i)}\big)^\top$, where I is the 2×2 identity matrix. By Lemma 3.1 with $A = -I$,

$$A\boldsymbol{\Sigma}A^\top = (-I)\,\boldsymbol{\Sigma}\,(-I)^\top = \boldsymbol{\Sigma},$$

and the mean is $A\boldsymbol{\mu} + \boldsymbol{b} = -\boldsymbol{\mu} + \big(y_1^{(n-j+1)},\, y_2^{(i)}\big)^\top = \boldsymbol{\mu}'$. □

Proposition 3.3 (Transformed coordinates: dominated cells)**.** For a dominated cell $C(i,j)$ with $i+j \geq n+1$, the transformation (67) yields

$$\bar{\boldsymbol{z}} = (\bar{z}_1, \bar{z}_2)^\top \sim \mathcal{N}\big(\boldsymbol{\mu}^-,\ \boldsymbol{\Sigma}\big), \qquad \boldsymbol{\mu}^- = \begin{pmatrix} \mu_1 - y_1^{(n-j+1)} \\ \mu_2 - y_2^{(i)} \end{pmatrix} = -\boldsymbol{\mu}'. \tag{69}$$

Again, $\mathrm{Cov}(\bar{\boldsymbol{z}}) = \boldsymbol{\Sigma}$.

Proof. Write $\bar{\boldsymbol{z}} = I\,\boldsymbol{y} - \big(y_1^{(n-j+1)},\, y_2^{(i)}\big)^\top$. By Lemma 3.1, $I\,\boldsymbol{\Sigma}\,I^\top = \boldsymbol{\Sigma}$ and the mean is $\boldsymbol{\mu} - \big(y_1^{(n-j+1)},\, y_2^{(i)}\big)^\top = \boldsymbol{\mu}^-$. □

As established in Remark 2.28, the covariance matrix $\boldsymbol{\Sigma}$ is invariant under both coordinate transformations, so when $\rho \neq 0$ the transformed variables remain correlated with the same coefficient ρ. Consequently, $\mathbb{E}[z_1 z_2] \neq \mathbb{E}[z_1]\,\mathbb{E}[z_2]$, and the distribution of the product $z_1 z_2$ cannot be obtained from the marginal distributions alone. This observation motivates the bivariate treatment developed in the remainder of this chapter.

Having established the distribution of the transformed coordinates, the joint density is recorded explicitly for later use.

Proposition 3.4 (Joint density of transformed coordinates)**.** The joint density of $\boldsymbol{z} = (z_1, z_2)^\top$ is

$$\varphi_{\boldsymbol{\mu}',\boldsymbol{\Sigma}}(z_1, z_2) \;=\; \frac{1}{2\pi\,\sigma_1\sigma_2\sqrt{1-\rho^2}}\;\exp\!\left(-\frac{Q(z_1,z_2)}{2(1-\rho^2)}\right), \tag{70}$$

where the quadratic form is

$$Q(z_1, z_2) \;=\; \frac{(z_1-\mu_1')^2}{\sigma_1^2} \;-\; \frac{2\rho\,(z_1-\mu_1')(z_2-\mu_2')}{\sigma_1\sigma_2} \;+\; \frac{(z_2-\mu_2')^2}{\sigma_2^2}\,. \tag{71}$$

Analogously, the joint density of $\bar{\boldsymbol{z}} = (\bar{z}_1, \bar{z}_2)^\top$ is $\varphi_{\boldsymbol{\mu}^-,\boldsymbol{\Sigma}}(\bar{z}_1, \bar{z}_2)$, obtained by replacing $\boldsymbol{\mu}'$ with $\boldsymbol{\mu}^-$ in (71).

Proof. Direct substitution of $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}$ into the bivariate Gaussian density (Definition 2.1), using the explicit inverse and determinant from Proposition 2.2. □

Finally, the truncation regions that restrict the transformed coordinates to a single cell are defined.

Definition 3.5 (Truncation bounds)**.** For a non-dominated cell $C(i,j)$ with $i+j \le n$, define

$$L_1 = y_1^{(n-j+1)} - u_1^{(i,j)}, \quad U_1 = y_1^{(n-j+1)} - l_1^{(i,j)}, \quad L_2 = y_2^{(i)} - u_2^{(i,j)}, \quad U_2 = y_2^{(i)} - l_2^{(i,j)}. \tag{72}$$

The truncation region is $\mathcal{R}^{(i,j)} = [L_1, U_1] \times [L_2, U_2]$.

For a dominated cell $C(i,j)$ with $i+j \ge n+1$, define

$$\bar{L}_1 = l_1^{(i,j)} - y_1^{(n-j+1)}, \quad \bar{U}_1 = u_1^{(i,j)} - y_1^{(n-j+1)}, \quad \bar{L}_2 = l_2^{(i,j)} - y_2^{(i)}, \quad \bar{U}_2 = u_2^{(i,j)} - y_2^{(i)}. \tag{73}$$

The truncation region is $\bar{\mathcal{R}}^{(i,j)} = [\bar{L}_1, \bar{U}_1] \times [\bar{L}_2, \bar{U}_2]$.

In both cases $L_k, \bar{L}_k \ge 0$ by construction (Section 2.5), so the product $z_1 z_2$ (or $\bar{z}_1\bar{z}_2$) is non-negative within the truncation region.

With the distributional results of this section in hand, the probability that $\boldsymbol{y}$ falls in a given cell, and the conditional density of $\Delta(\boldsymbol{y})$ within that cell, can now be expressed in terms of the bivariate Gaussian density (70). These expressions are developed in Sections 3.3 and 3.4.

3.3 Cell Probabilities Under Correlation

The distribution of the generalized HVI is obtained by marginalizing over all cells of the partition (Section 2.5). Each cell's contribution is weighted by the probability that $\boldsymbol{y}$ falls in that cell. Under independent objectives this probability factors as a product of two univariate CDF differences[27, Eq. 3]. It is now shown that, under correlation, the cell probability becomes a bivariate normal rectangle probability that does not factor unless $\rho = 0$.

Recall the cell probability from Definition 2.29 (Eq. (52)).

To evaluate (52) in closed form, the bounds are standardized and the result is expressed in terms of the standard bivariate normal CDF.

Theorem 3.6 (Cell probability via bivariate CDF)**.** Define the standardized bounds for a non-dominated cell:

$$\tilde{L}_k = \frac{L_k - \mu'_k}{\sigma_k}, \qquad \tilde{U}_k = \frac{U_k - \mu'_k}{\sigma_k}, \qquad k = 1, 2. \tag{74}$$

For a dominated cell, replace L_k, U_k, μ'_k by $\bar{L}_k, \bar{U}_k, \mu_k^-$ respectively. Then the cell probability is

$$P_{\text{cell}}^{(i,j)} \;=\; \Phi_2\big(\tilde{U}_1, \tilde{U}_2;\, \rho\big) \;-\; \Phi_2\big(\tilde{U}_1, \tilde{L}_2;\, \rho\big) \;-\; \Phi_2\big(\tilde{L}_1, \tilde{U}_2;\, \rho\big) \;+\; \Phi_2\big(\tilde{L}_1, \tilde{L}_2;\, \rho\big), \tag{75}$$

where $\Phi_2(a, b;\, \rho) = \Pr(Z_1 \le a,\, Z_2 \le b)$ for $(Z_1, Z_2) \sim \mathcal{N}\big(\mathbf{0},\, \Sigma_\rho\big)$ with $\Sigma_\rho = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$.

Proof. Standardize via $\tilde{z}_k = (z_k - \mu'_k)/\sigma_k$ (or $(\bar{z}_k - \mu_k^-)/\sigma_k$ for dominated cells). Then $(\tilde{z}_1, \tilde{z}_2) \sim \mathcal{N}(\mathbf{0}, \Sigma_\rho)$, and the inclusion–exclusion identity

$$\Pr(a_1 \le X \le b_1,\; a_2 \le Y \le b_2) \;=\; F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2)$$

yields (75). □

Remark 3.7 (Reduction to the independent case). When $\rho = 0$, the standard bivariate normal CDF factors: $\Phi_2(a, b;\, 0) = \Phi(a)\,\Phi(b)$. Substituting into (75) gives

$$P_{\text{cell}}^{(i,j)} \;=\; \big[\Phi(\tilde{U}_1) - \Phi(\tilde{L}_1)\big] \;\cdot\; \big[\Phi(\tilde{U}_2) - \Phi(\tilde{L}_2)\big],$$

recovering the product form of Wang et al. [27, Eq. 3]. This factorization holds for both non-dominated and dominated cells. When $\rho \neq 0$, the four bivariate CDF evaluations in (75) cannot be simplified further, and the probability of falling in a cell depends on the joint behavior of both objectives rather than on each objective independently.

Remark 3.8 (Numerical evaluation). The standard bivariate normal CDF $\Phi_2(a, b;\, \rho)$ admits efficient numerical evaluation via the algorithm of Drezner and Wesolowsky [11], which reduces the computation to a one-dimensional integral evaluated by Gauss-Legendre quadrature with guaranteed absolute error below 10^{-15}. Implementations are available in standard numerical libraries (e.g., `scipy.stats.mvn` in Python, `mvtnorm` in R). Each cell probability therefore costs $O(1)$ to evaluate, and the total cost across all $(n+1)^2$ cells remains $O(n^2)$, the same order as in the independent case.

The within-cell distribution also requires the conditional density of (z_1, z_2) restricted to the truncation region is recorded here for later use.

Definition 3.9 (Truncated bivariate normal density)**.** The conditional density of (z_1, z_2) given $\boldsymbol{z} \in \mathcal{R}^{(i,j)}$ (non-dominated cell) is

$$f_{z_1,z_2}^{\text{trunc}}(z_1, z_2) \;=\; \frac{\varphi_{\boldsymbol{\mu}', \boldsymbol{\Sigma}}(z_1, z_2)}{P_{\text{cell}}^{(i,j)}}\; \mathbf{1}_{(z_1,z_2) \in \mathcal{R}^{(i,j)}}. \tag{76}$$

For a dominated cell, replace $\boldsymbol{\mu}'$ by $\boldsymbol{\mu}^-$ and $\mathcal{R}^{(i,j)}$ by $\bar{\mathcal{R}}^{(i,j)}$.

3.4 Product Density Under Correlation

Within each cell of the partition, the generalized HVI is expressed as a product of the transformed coordinates plus a cell-dependent offset (Section 2.4.2). The probability distribution of HVI within a cell is therefore determined by the distribution of the product $Z = z_1 z_2$ (or $\bar{Z} = \bar{z}_1 \bar{z}_2$) under the truncated bivariate Gaussian. The derivation begins with a general change-of-variables result, then specializes to each cell type.

By Lemma 2.30, the product $Z = XY$ has density given by (53). This is now specialized to each cell type.

Theorem 3.10 (Product density: non-dominated cells)**.** Let (z_1, z_2) follow the truncated bivariate normal on $\mathcal{R}^{(i,j)} = [L_1, U_1] \times [L_2, U_2]$ with $0 \le L_k \le U_k$. Define $Z = z_1 z_2$. Then

$$f_Z^{\text{trunc}}(z) \;=\; \frac{1}{P_{\text{cell}}^{(i,j)}} \int_{\alpha(z)}^{\beta(z)} \varphi_{\boldsymbol{\mu}',\boldsymbol{\Sigma}}\Big(x,\ \tfrac{z}{x}\Big)\, \frac{1}{x}\, \mathrm{d}x, \qquad z \in [L_1 L_2,\ U_1 U_2], \tag{77}$$

and $f_Z^{\text{trunc}}(z) = 0$ otherwise, with integration bounds

$$\alpha(z) = \max\Big(L_1,\ \tfrac{z}{U_2}\Big), \qquad \beta(z) = \begin{cases} \min\big(U_1,\ \tfrac{z}{L_2}\big), & L_2 > 0, \\ U_1, & L_2 = 0. \end{cases} \tag{78}$$

Proof. Since $L_1, L_2 \ge 0$ (Proposition 2.25), we have $z_1, z_2 \ge 0$ almost surely under the truncation, so $Z \ge 0$ and $|x| = x$. Apply Lemma 2.30 to the truncated density (76) and enforce the constraints $x \in [L_1, U_1]$ and $z/x \in [L_2, U_2]$ simultaneously to obtain the bounds (78). □

Theorem 3.11 (Product density: dominated cells)**.** Let $(\bar{z}_1, \bar{z}_2)$ follow the truncated bivariate normal on $\bar{\mathcal{R}}^{(i,j)} = [\bar{L}_1, \bar{U}_1] \times [\bar{L}_2, \bar{U}_2]$ with $0 \le \bar{L}_k \le \bar{U}_k$. Define $\bar{Z} = \bar{z}_1 \bar{z}_2$. Then

$$f_{\bar{Z}}^{\text{trunc}}(\bar{z}) \;=\; \frac{1}{P_{\text{cell}}^{(i,j)}} \int_{\bar{\alpha}(\bar{z})}^{\bar{\beta}(\bar{z})} \varphi_{\boldsymbol{\mu}^-,\boldsymbol{\Sigma}}\Big(x,\ \tfrac{\bar{z}}{x}\Big)\, \frac{1}{x}\, \mathrm{d}x, \qquad \bar{z} \in [\bar{L}_1 \bar{L}_2,\ \bar{U}_1 \bar{U}_2], \tag{79}$$

and zero otherwise, with bounds

$$\bar{\alpha}(\bar{z}) = \max\Big(\bar{L}_1,\ \tfrac{\bar{z}}{\bar{U}_2}\Big), \qquad \bar{\beta}(\bar{z}) = \begin{cases} \min\big(\bar{U}_1,\ \tfrac{\bar{z}}{\bar{L}_2}\big), & \bar{L}_2 > 0, \\ \bar{U}_1, & \bar{L}_2 = 0. \end{cases} \tag{80}$$

Proof. Identical to Theorem 3.10 with $(\boldsymbol{\mu}', L_k, U_k)$ replaced by $(\boldsymbol{\mu}^-, \bar{L}_k, \bar{U}_k)$. □

The integrands in (77) and (79) involve the bivariate Gaussian density evaluated along the hyperbola $z_2 = z/z_1$. The next result reveals the algebraic structure of the exponent under this substitution, which is the key to understanding why correlation fundamentally changes the integral.

By Proposition 2.33, the quadratic form evaluated at $(x,\ z/x)$ takes the Laurent polynomial form (58) with coefficients (59)–(63).

Remark 3.12 (Effect of correlation on the exponent). Setting $\rho = 0$ eliminates all coupling between x and z in the exponent:

$$a_1 = -\frac{2\nu_1}{\sigma_1^2}, \qquad a_0 = \frac{\nu_1^2}{\sigma_1^2} + \frac{\nu_2^2}{\sigma_2^2},$$
$$b_1 = -\frac{2\nu_2\, z}{\sigma_2^2}.$$

In this case, the exponent separates as $Q(x, z/x) = g_1(x) + g_2(z/x)$, where g_1 depends only on x and g_2 only on z/x. The Gaussian density then factors: $\varphi_{\boldsymbol{\nu},\boldsymbol{\Sigma}}(x, z/x) = \varphi_{\nu_1,\sigma_1}(x) \cdot \varphi_{\nu_2,\sigma_2}(z/x)$, and the product density integral (77) reduces to the one-dimensional quadrature formula of Wang et al. [27, Eq. 4].

When $\rho \neq 0$, the ρ-dependent terms in a_1, a_0, and b_1 prevent this separation. The exponent is a genuinely two-variable Laurent polynomial in x and x^{-1} with coefficients that depend on both z and ρ, and the integral (77) must be evaluated as a single one-dimensional quadrature over the full bivariate density. This non-separability is the analytical consequence of introducing correlation between the objectives.

Remark 3.13 (Identical structure for dominated cells). Since the only difference between the non-dominated and dominated product densities is the shifted mean ($\boldsymbol{\mu}'$ versus $\boldsymbol{\mu}^-$), the Laurent polynomial structure (58)-(63) applies to both cases with the appropriate substitution $\boldsymbol{\nu} = \boldsymbol{\mu}'$ or $\boldsymbol{\nu} = \boldsymbol{\mu}^-$.

The product density is now connected to the conditional distribution of the generalized HVI within each cell.

Theorem 3.14 (Conditional PDF of generalized HVI: non-dominated cells)**.** For a non-dominated cell $C(i,j)$ with $i + j \leq n$, the conditional density of $\Delta(\boldsymbol{y})\big|_{C(i,j)} = z_1 z_2 + \gamma^{(i,j)}$, given $\boldsymbol{y} \in C(i,j)$, is

$$\mathrm{PDF}_{\Delta}^{(i,j)}(\delta) \;=\; f_Z^{\mathrm{trunc}}\big(\delta - \gamma^{(i,j)}\big), \qquad \delta \in \big[\Delta^+(u^{(i,j)}),\ \Delta^+(l^{(i,j)})\big], \tag{81}$$

where f_Z^{trunc} is from Theorem 3.10.

Proof. Since $\Delta = Z + \gamma^{(i,j)}$ with $Z = z_1 z_2$ and $\gamma^{(i,j)}$ a constant, the density of Δ at δ equals the density of Z at $p = \delta - \gamma^{(i,j)}$. □

Theorem 3.15 (Conditional PDF of generalized HVI: dominated cells)**.** For a dominated cell $C(i,j)$ with $i + j \geq n + 1$, where $\Delta(\boldsymbol{y})\big|_{C(i,j)} = -\bar{z}_1 \bar{z}_2 + \gamma^{-(i,j)}$, the conditional density satisfies

$$\mathrm{PDF}_{\Delta^-}^{(i,j)}(\delta) \;=\; f_{\bar{Z}}^{\mathrm{trunc}}\big(\gamma^{-(i,j)} - \delta\big), \qquad \delta \in \big[\Delta^-(u^{(i,j)}),\ \Delta^-(l^{(i,j)})\big]. \tag{82}$$

Proof. Since $\Delta^- = -\bar{Z} + \gamma^{-(i,j)}$ with $\bar{Z} \geq 0$, we have $\Pr(\Delta^- \leq \delta) = \Pr(\bar{Z} \geq \gamma^{-(i,j)} - \delta)$. Differentiating gives $f_{\Delta^-}^{(i,j)}(\delta) = f_{\bar{Z}}^{\mathrm{trunc}}(\gamma^{-(i,j)} - \delta)$. □

As noted in Remark 2.20, the dominated-cell density can be related to the non-dominated formula via coordinate inversion: $\mathrm{PDF}_{\Delta^-}^{(i,j)}(\delta) = \mathrm{PDF}_{\Delta^+}^{(i,j)}(-\delta)$. This equivalence extends without modification to the correlated setting because $\boldsymbol{\Sigma}$ is preserved under the negation map (Propositions 3.2 and 3.3).

Theorem 3.16 (Conditional CDF of generalized HVI)**.** For a non-dominated cell $C(i,j)$ with $i+j \le n$, the conditional CDF is

$$F_{\Delta}^{(i,j)}(\delta) \;=\; \Pr\big(Z \le p \;\big|\; \boldsymbol{z} \in \mathcal{R}^{(i,j)}\big), \qquad p = \delta - \gamma^{(i,j)}, \tag{83}$$

which can be expressed as

$$F_{\Delta}^{(i,j)}(\delta) \;=\; \frac{1}{P_{\text{cell}}^{(i,j)}} \iint_{\mathcal{R}^{(i,j)}(p)} \varphi_{\boldsymbol{\mu}',\boldsymbol{\Sigma}}(z_1,z_2)\,\mathrm{d}z_1\,\mathrm{d}z_2, \tag{84}$$

where $\mathcal{R}^{(i,j)}(p) = \{(z_1,z_2) \in \mathcal{R}^{(i,j)} : z_1 z_2 \le p\}$.

For a dominated cell $C(i,j)$ with $i+j \ge n+1$:

$$F_{\Delta}^{(i,j)}(\delta) \;=\; 1 - \Pr\big(\bar{Z} < \gamma^{-(i,j)} - \delta \;\big|\; \bar{\boldsymbol{z}} \in \bar{\mathcal{R}}^{(i,j)}\big), \tag{85}$$

or equivalently, via the coordinate inversion of Remark 2.20: $F_{\Delta^-}^{(i,j)}(\delta) = 1 - F_{\Delta^+}^{(i,j)}(-\delta)$.

Proof. The non-dominated case follows directly from $\Delta = Z + \gamma^{(i,j)}$. For the dominated case, $\Delta^- = -\bar{Z} + \gamma^{-(i,j)}$, so $\Pr(\Delta^- \le \delta) = \Pr(\bar{Z} \ge \gamma^{-(i,j)} - \delta) = 1 - \Pr(\bar{Z} < \gamma^{-(i,j)} - \delta)$. The inversion equivalence follows from Remark 2.20. □

Remark 3.17 (Geometry of the sublevel set). The boundary $z_1 z_2 = p$ (or $\bar{z}_1 \bar{z}_2 = p$) is a rectangular hyperbola in the first quadrant. As δ varies across the cell's HVI range, the hyperbola sweeps across the truncation rectangle, and the conditional CDF increases from 0 to 1. In the independent case, the integrand along this hyperbola factors into a product of univariate densities; under correlation, the full bivariate density (70) must be retained, and the integral is evaluated numerically using, e.g., Gauss-Kronrod quadrature (21-point rule with up to 50 sub-intervals, following Wang et al. [27]).

3.5 Full Marginal Distribution

The results of Subsections 3.3 and 3.4 are now assembled into the unconditional distribution of the generalized hypervolume improvement under correlated Gaussian objectives.

Theorem 3.18 (Marginal distribution of generalized HVI)**.** The unconditional CDF of the generalized hypervolume improvement under the correlated posterior (65) is

$$F_{\Delta}(\delta) \;=\; \sum_{i,j=0}^{n} P_{\text{cell}}^{(i,j)}\; F_{\Delta}^{(i,j)}(\delta), \tag{86}$$

where $F_{\Delta}^{(i,j)}$ is the conditional CDF from Theorem 3.16 (using the non-dominated formula (83) when $i+j \le n$ and the dominated formula (85) when $i+j \ge n+1$).

The unconditional PDF is

$$f_{\Delta}(\delta) \;=\; \sum_{i,j=0}^{n} P_{\text{cell}}^{(i,j)}\; f_{\Delta}^{(i,j)}(\delta), \tag{87}$$

where $f_{\Delta}^{(i,j)}$ is from Theorem 3.14 (non-dominated) or Theorem 3.15 (dominated).

Proof. By the law of total probability:

$$F_\Delta(\delta) \;=\; \sum_{i,j=0}^{n} P_{\text{cell}}^{(i,j)}\;\Pr\big(\Delta(\boldsymbol{y}) \le \delta \;\big|\; \boldsymbol{y} \in C(i,j)\big).$$

Each conditional CDF is continuous (the generalized HVI is continuous within each cell by the smoothness of the product $z_1 z_2$ and the bivariate Gaussian density), so the sum is a continuous mixture. Unlike the non-generalized case, there is no point mass at $\delta = 0$: dominated cells contribute continuous negative-valued densities rather than concentrating all their probability at zero. The PDF follows by differentiating (86). □

Remark 3.19 (Reduction to non-generalized HVI). If one restricts attention to the standard (non-negative) HVI Δ^+ by setting $\Delta(\boldsymbol{y}) = 0$ for dominated points, the marginal CDF becomes

$$F_{\Delta^+}(\delta) \;=\; \big(1 - P_{\text{nd}}\big)\,\mathbf{1}[\delta \ge 0] \;+\; \sum_{\substack{i,j=0 \\ i+j\le n}}^{n} P_{\text{cell}}^{(i,j)}\; F_\Delta^{(i,j)}(\delta), \tag{88}$$

where $P_{\text{nd}} = \sum_{i+j\le n} P_{\text{cell}}^{(i,j)}$ is the total probability of landing in a non-dominated cell. This recovers the point-mass structure: all dominated cells contribute a single atom at $\delta = 0$ with total weight $1 - P_{\text{nd}}$.

The central result of this chapter is that the correlated distribution contains the independent case as a strict special case.

Theorem 3.20 (Consistency with Wang et al. (2024))**.** When $\rho = 0$, the correlated formulas in Theorems 3.6, 3.10, 3.11, 3.14, 3.15, and 3.18 reduce to the independent-objective expressions of Wang et al. [27].

Proof. We verify each component.

(i) Cell probability. When $\rho = 0$, $\Phi_2(a, b; 0) = \Phi(a)\,\Phi(b)$. Substituting into Theorem 3.6:

$$P_{\text{cell}}^{(i,j)} \;=\; \big[\Phi(\tilde{U}_1) - \Phi(\tilde{L}_1)\big] \;\cdot\; \big[\Phi(\tilde{U}_2) - \Phi(\tilde{L}_2)\big],$$

recovering the product form of Wang et al. [27, Eq. 3]. This holds for both non-dominated and dominated cells. *(ii) Joint density factorization.* When $\rho = 0$, the quadratic form (71) separates and the bivariate density factors: $\varphi_{\boldsymbol{\nu},\boldsymbol{\Sigma}}(z_1, z_2) = \varphi_{\nu_1,\sigma_1}(z_1) \cdot \varphi_{\nu_2,\sigma_2}(z_2)$.

(iii) Laurent coefficients. Setting $\rho = 0$ in Proposition 2.33: $a_1 \to -2\nu_1/\sigma_1^2$; the ρ-dependent parts of a_0 and b_1 vanish; and the exponent separates as $Q(x, z/x) = g_1(x) + g_2(z/x)$ (Remark 3.12).

(iv) Product density integral. With the factorized density, the integrand in (77) (and (79)) becomes $\varphi_{\nu_1,\sigma_1}(x) \cdot \varphi_{\nu_2,\sigma_2}(z/x) \cdot x^{-1}$, the one-dimensional quadrature formula of Wang et al. [27, Eq. 4].

(v) Dominated cells. The coordinate inversion of Remark 2.20 reduces to the inversion trick of Wang et al. [27]: $\text{PDF}_{\Delta^-}^{(i,j)}(\delta) = \text{PDF}_{\Delta^+}^{(i,j)}(-\delta)$ with inverted quantities, and $\text{CDF}_{\Delta^-}^{(i,j)}(\delta) = 1 - \text{CDF}_{\Delta^+}^{(i,j)}(-\delta)$, exactly matching their formulas.

(vi) Marginal distribution. The continuous mixture over all cells in Theorem 3.18 is consistent with the marginal distribution of Wang et al., which sums over both non-dominated and dominated

cells. When restricted to Δ^+ only (Remark 3.19), the point-mass treatment of dominated cells is also recovered.

Thus the correlated theory with generalized HVI is a strict generalization of the independent case. □

Summary. This section has derived the exact CDF and PDF of the generalized hypervolume improvement under correlated bivariate Gaussian posteriors. The distribution is a continuous mixture over all $(n+1)^2$ cells of the partition, where each cell contributes a density obtained from the product of truncated correlated Gaussian coordinates. The key structural differences from the independent case are twofold: cell probabilities require bivariate normal CDF evaluations (Theorem 3.6), and the within-cell product density involves a non-separable Laurent polynomial exponent (Proposition 2.33). The entire framework reduces exactly to Wang et al. [27] when $\rho = 0$ (Theorem 3.20).

4 Extension II: Constrained Hypervolume Improvement

This section develops the second main contribution of the thesis: the probability distribution of the hypervolume improvement under black-box inequality constraints. The presentation builds on the correlated HVI distribution derived in Section 3 and on the Gaussian conditioning and cell-partition tools established in Chapter 2.

The central idea is simple. Rather than treating the acquisition value and constraint satisfaction as separate, independently combined factors (the approach of Gardner et al. [14]), a single random variable is defined:

$$\Delta_C \;=\; \Delta(\boldsymbol{y}) \;\cdot\; \mathbf{1}\big[\boldsymbol{g} \preceq \mathbf{0}\big] \tag{89}$$

that equals the generalized HVI when all constraints are satisfied and zero otherwise, and derive its exact distribution. Everything in this section flows from (89).

4.1 Constrained Random Variable Definition

We consider a bi-objective optimization problem with J black-box inequality constraints:

$$\min_{\boldsymbol{x}\in\mathcal{X}\subseteq\mathbb{R}^d} \boldsymbol{f}(\boldsymbol{x}) = \big(f_1(\boldsymbol{x}),\, f_2(\boldsymbol{x})\big)^\top \qquad \text{subject to} \quad c_j(\boldsymbol{x}) \le 0,\;\; j \in [1..J], \tag{90}$$

where the objectives f_i and constraints c_j are expensive black-box functions, each modeled with a Gaussian process prior. At a candidate point $\boldsymbol{x}$, the posterior predictive distribution for each constraint is (Section 2.1)

$$g_j \;:=\; c_j(\boldsymbol{x}) \mid \mathcal{D} \;\sim\; \mathcal{N}\big(\mu_{c_j},\, \sigma^2_{c_j}\big), \qquad j \in [1..J]. \tag{91}$$

The relationship between the objective posteriors, whether independent or correlated, is treated in Section 3; the constraint extension developed here accommodates both cases.

Remark 4.1 (Shared observation set). Throughout this section we assume that all objectives and constraints share the same observation locations $\boldsymbol{X} \subset \mathcal{X}$, so that the joint posterior (102) is well-defined at every candidate point. In practice, constraints may sometimes be observed at different locations than the objectives (e.g., when safety tests are conducted separately); extending the framework to heterogeneous observation sets is straightforward via the standard GP predictive equations but is not pursued here.

4.1.1 Feasibility indicator

Definition 4.2 (Feasibility indicator)**.** The feasibility indicator at a point $\boldsymbol{x}$ is

$$I_{\text{feas}}(\boldsymbol{x}) \;:=\; \mathbf{1}\Big[\bigcap_{j=1}^{J}\{c_j(\boldsymbol{x}) \le 0\}\Big] \;=\; \prod_{j=1}^{J} \mathbf{1}\big[g_j \le 0\big]. \tag{92}$$

4.1.2 Constrained generalized HVI

Definition 4.3 (Constrained generalized hypervolume improvement)**.** Given the generalized HVI $\Delta(\boldsymbol{y})$ (Section 2.4.2) and the feasibility indicator I_{feas}, the *constrained generalized hypervolume*

improvement is

$$\Delta_C(\boldsymbol{y}, \boldsymbol{g}) \;:=\; I_{\text{feas}} \cdot \Delta(\boldsymbol{y}) \;=\; \Delta(\boldsymbol{y}) \;\cdot\; \prod_{j=1}^{J} \mathbf{1}\big[g_j \le 0\big], \tag{93}$$

where $\boldsymbol{y} = (y_1, y_2)^\top$ are the objective values and $\boldsymbol{g} = (g_1, \ldots, g_J)^\top$ are the constraint values at $\boldsymbol{x}$.

Remark 4.4 (Recovery of the unconstrained case). When $J = 0$ (no constraints), $\Delta_C \equiv \Delta$, recovering the unconstrained setting of Wang et al. [27] and Section 3.

Remark 4.5 (Two-level priority structure). For an infeasible point ($I_{\text{feas}} = 0$), the constrained HVI is zero regardless of whether $\boldsymbol{y}$ is dominated or non-dominated. The negative HVI Δ^- only takes effect for feasible dominated points, penalizing feasible evaluations that fail to improve $\mathcal{P}$. This creates a natural two-level structure: feasibility is the first priority; objective improvement is the second.

4.1.3 Mixed distribution structure

Under the GP posterior at a candidate $\boldsymbol{x}$, both $\boldsymbol{y}$ and $\boldsymbol{g}$ are random variables, making Δ_C a real-valued random variable.

Proposition 4.6 (Mixed distribution of Δ_C)**.** Δ_C is a mixed random variable with:

1. A *point mass at zero* of magnitude $\Pr(\Delta_C = 0)$, arising from infeasibility.
2. A *continuous component* inherited from the continuous distribution of $\Delta(\boldsymbol{y})$ restricted to feasible outcomes.

Proof. Under continuous GP posteriors, the event $\{\Delta(\boldsymbol{y}) = 0,\ I_{\text{feas}} = 1\}$ (a feasible point landing exactly on the attainment boundary) has probability zero almost surely. Therefore the atom at zero arises entirely from infeasibility: $\Pr(\Delta_C = 0) = 1 - \Pr(I_{\text{feas}} = 1)$. Away from zero, Δ_C inherits the continuous density of $\Delta(\boldsymbol{y})$ conditional on all constraints being satisfied. □

4.2 Independent Constraint Case

The distribution of Δ_C is first derived in the simplest setting: all constraints are independent of each other and of the objectives.

Assumption 4.7 (Independence)**.** The constraint GPs are mutually independent and independent of both objective GPs:

$$y_1 \;\perp\; y_2 \;\perp\; g_1 \;\perp\; \cdots \;\perp\; g_J \;\Big|\; \mathcal{D}. \tag{94}$$

Under Assumption 4.7, the feasibility indicator is independent of the objectives, and the joint probability of feasibility factors as a product of marginal probabilities.

Definition 4.8 (Probability of feasibility)**.** The posterior probability of feasibility for the j-th constraint is

$$\text{PoF}_j(\boldsymbol{x}) \;:=\; \Pr(g_j \le 0 \mid \mathcal{D}, \boldsymbol{x}) \;=\; \Phi\!\left(\frac{-\mu_{c_j}}{\sigma_{c_j}}\right). \tag{95}$$

Under Assumption 4.7, the joint probability of feasibility is

$$\mathrm{PoF}(\boldsymbol{x}) \;=\; \prod_{j=1}^{J} \mathrm{PoF}_j(\boldsymbol{x}). \tag{96}$$

4.2.1 CDF via the law of total probability

The distribution of Δ_C follows by conditioning on feasibility.

Theorem 4.9 (CDF of constrained HVI: independent constraints)**.** Under Assumption 4.7, the CDF of the constrained generalized HVI is, for all $\delta \in \mathbb{R}$:

$$F_{\Delta_C}(\delta) \;=\; \big(1 - \mathrm{PoF}(\boldsymbol{x})\big) \cdot \mathbf{1}[\delta \geq 0] \;+\; \mathrm{PoF}(\boldsymbol{x}) \cdot F_{\Delta}(\delta), \tag{97}$$

where $F_{\Delta}(\delta)$ is the unconstrained generalized HVI CDF from Theorem 3.18.

Proof. Since $\Delta_C = \Delta \cdot \mathbf{1}[\boldsymbol{g} \preceq \mathbf{0}]$ with $\Delta \perp \mathbf{1}[\boldsymbol{g} \preceq \mathbf{0}]$ (by Assumption 4.7), condition on feasibility via the law of total probability:

$$F_{\Delta_C}(\delta) = \Pr(\Delta_C \leq \delta \mid \boldsymbol{g} \preceq \mathbf{0})\, \mathrm{PoF}(\boldsymbol{x}) \;+\; \Pr(\Delta_C \leq \delta \mid \boldsymbol{g} \not\preceq \mathbf{0})\, \big(1 - \mathrm{PoF}(\boldsymbol{x})\big). \tag{98}$$

When $\boldsymbol{g} \not\preceq \mathbf{0}$ (infeasible), $\Delta_C = 0$, so $\Pr(\Delta_C \leq \delta \mid \boldsymbol{g} \not\preceq \mathbf{0}) = \mathbf{1}[\delta \geq 0]$. When $\boldsymbol{g} \preceq \mathbf{0}$ (feasible), $\Delta_C = \Delta(\boldsymbol{y})$, and by independence the conditioning does not alter the distribution of $\boldsymbol{y}$: $\Pr(\Delta_C \leq \delta \mid \boldsymbol{g} \preceq \mathbf{0}) = F_{\Delta}(\delta)$. Substituting yields (97). □

4.2.2 PDF and mixture decomposition

Theorem 4.10 (PDF of constrained HVI: independent constraints)**.** The constrained HVI has a mixed distribution with:

1. A point mass at $\delta = 0$ of magnitude $1 - \mathrm{PoF}(\boldsymbol{x})$.
2. A continuous density for $\delta \neq 0$:

$$f_{\Delta_C}(\delta) \;=\; \mathrm{PoF}(\boldsymbol{x}) \;\cdot\; f_{\Delta}(\delta), \qquad \delta \neq 0, \tag{99}$$

where f_{Δ} is the unconstrained generalized HVI density from Theorem 3.18, evaluated via the cell partition (Section 2.5).

Proof. Differentiate (97) with respect to δ away from $\delta = 0$. The indicator $\mathbf{1}[\delta \geq 0]$ contributes a Dirac delta at zero for the point mass; the smooth term differentiates to $\mathrm{PoF}(\boldsymbol{x}) \cdot f_{\Delta}(\delta)$. □

4.2.3 Cell-level conditional density

Expanding the cell-partition structure, the contribution from each cell is simply rescaled by the probability of feasibility. For a non-dominated cell $C(i,j)$ with $i + j \leq n$, define $z_1 = y_1^{(n-j+1)} - y_1$ and $z_2 = y_2^{(i)} - y_2$ as in Section 3.4, with $p = \delta - \gamma^{(i,j)}$. The cell-level density of the constrained HVI is

$$\mathrm{PDF}_{\Delta_C}^{(i,j)}(\delta) \;=\; \mathrm{PoF}(\boldsymbol{x}) \;\cdot\; \mathrm{PDF}_{\Delta}^{(i,j)}(\delta), \tag{100}$$

where $\mathrm{PDF}_{\Delta}^{(i,j)}$ is the unconstrained cell-level density from Theorem 3.14 (if $\rho \neq 0$, using the correlated bivariate formula) or from the factorized Wang et al. formula (if $\rho = 0$). For dominated cells ($i + j > n$), the coordinate-inversion trick (Remark 2.20) applies identically, with the overall density rescaled by $\mathrm{PoF}(\boldsymbol{x})$.

Remark 4.11 (Product weighting is exact under independence). Equations (97) and (99) show that, under Assumption 4.7, the constrained HVI distribution is obtained by multiplying the unconstrained distribution by a scalar feasibility weight. This is precisely the Gardner et al. [14] product weighting scheme:

$$\alpha_{\text{constrained}}(\boldsymbol{x}) \;=\; \alpha(\boldsymbol{x}) \;\cdot\; \prod_{j=1}^{J} \mathrm{PoF}_j(\boldsymbol{x}). \tag{101}$$

This derivation shows that the product approach is *exact* when constraints are independent of objectives, not an ad hoc heuristic, but a consequence of the law of total probability. Conversely, when constraints and objectives are correlated, the product weighting is no longer exact: conditioning on feasibility modifies the objective distribution, and the unconstrained HVI density f_{Δ} must be replaced by the conditional density $f_{\Delta|\boldsymbol{g}\preceq\mathbf{0}}$. This is the subject of Subsection 4.3.

Remark 4.12 (Computational cost). Computing the constrained distribution under independence adds negligible cost to the unconstrained evaluation: the only additional work is J univariate Gaussian CDF evaluations for $\mathrm{PoF}_j(\boldsymbol{x})$. The total complexity per candidate point is therefore $O\big(\tau^{-1}(n+1)^2 + J\big)$, where $\tau^{-1}(n+1)^2$ is the cost of the unconstrained ε-PoHVI evaluation from Wang et al. [27].

4.3 Distribution Under a Single Correlated Constraint

The independence assumption of Section 4.2 is now relaxed. Suppose a multi-output Gaussian process, for instance a convolution process [1] or a linear model of coregionalization [6], is used to jointly model the objectives and the constraint, yielding a joint posterior at a candidate point $\boldsymbol{x}$:

$$\begin{pmatrix} y_1 \\ y_2 \\ g \end{pmatrix} \,\Big|\, \mathcal{D}, \boldsymbol{x} \;\sim\; \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_c \end{pmatrix},\; \boldsymbol{\Sigma} \right), \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \rho_1\sigma_1\sigma_c \\ \sigma_{12} & \sigma_2^2 & \rho_2\sigma_2\sigma_c \\ \rho_1\sigma_1\sigma_c & \rho_2\sigma_2\sigma_c & \sigma_c^2 \end{pmatrix}, \tag{102}$$

where $\rho_i = \mathrm{Corr}(y_i, g)$ captures the objective-constraint correlation, and $\sigma_{12} = \mathrm{Cov}(y_1, y_2)$ is the objective-objective covariance. The latter may be zero when independent objective GPs are used, or nonzero when the objectives are modeled with a correlated GP as developed in Chapter 3.

Remark 4.13 (Generality of the formulation). Equation (102) accommodates three scenarios:

1. *Independent objectives, independent constraint* ($\sigma_{12} = 0$, $\rho_1 = \rho_2 = 0$): reduces to the case of Section 4.2.

2. *Independent objectives, correlated constraint* ($\sigma_{12} = 0$, $\rho_1 \neq 0$ or $\rho_2 \neq 0$): the objectives are marginally independent but both may correlate with g.

3. *Correlated objectives, correlated constraint* ($\sigma_{12} \neq 0$, $\rho_i \neq 0$): the fully general case.

4.3.1 Conditioning on the Constraint Induces Objective Dependence

The central observation of this section is that conditioning on the constraint value g, whether on the event $g = t$ or on the event $g \leq 0$, generically modifies the dependence structure between y_1 and y_2. This is a direct consequence of the Gaussian conditioning formula stated in Section 2.3.

Lemma 4.14 (Conditional distribution of objectives given constraint)**.** Let $(y_1, y_2, g)^\top$ be jointly Gaussian as in (102). By the Gaussian conditioning formula (Section 2.3), the conditional distribution of (y_1, y_2) given $g = t$ is:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \Bigg| g = t, \mathcal{D} \;\sim\; \mathcal{N}\big(\boldsymbol{\mu}_{y|g}(t),\; \boldsymbol{\Sigma}_{y|g}\big), \tag{103}$$

where:

$$\boldsymbol{\mu}_{y|g}(t) = \begin{pmatrix} \mu_1 + \rho_1 \dfrac{\sigma_1}{\sigma_c}(t - \mu_c) \\ \mu_2 + \rho_2 \dfrac{\sigma_2}{\sigma_c}(t - \mu_c) \end{pmatrix}, \tag{104}$$

$$\boldsymbol{\Sigma}_{y|g} = \begin{pmatrix} \sigma_1^2(1 - \rho_1^2) & \sigma_{12} - \rho_1\rho_2\sigma_1\sigma_2 \\ \sigma_{12} - \rho_1\rho_2\sigma_1\sigma_2 & \sigma_2^2(1 - \rho_2^2) \end{pmatrix}. \tag{105}$$

The conditional mean $\boldsymbol{\mu}_{y|g}(t)$ is affine in t, while $\boldsymbol{\Sigma}_{y|g}$ is independent of t.

Proof. Partition the joint distribution (102) as $(\boldsymbol{y}, g)$ with $\boldsymbol{\Sigma}_{yg} = (\rho_1\sigma_1\sigma_c,\; \rho_2\sigma_2\sigma_c)^\top$ and $\Sigma_{gg} = \sigma_c^2$. By the conditioning formula of Section 2.3: $\boldsymbol{\mu}_{y|g}(t) = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yg}\Sigma_{gg}^{-1}(t - \mu_c)$ and $\boldsymbol{\Sigma}_{y|g} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yg}\Sigma_{gg}^{-1}\boldsymbol{\Sigma}_{yg}^\top$. Substituting yields (104)-(105). □

Remark 4.15 (Positive definiteness and conditional correlation). Since $\boldsymbol{\Sigma}_{y|g}$ is the Schur complement of Σ_{gg} in the positive definite joint covariance (102), it is itself positive definite whenever $|\rho_1| < 1$ and $|\rho_2| < 1$. In particular, the conditional variances $\sigma_i^2(1 - \rho_i^2) > 0$ and the conditional correlation coefficient

$$\rho_{y|g} \;=\; \frac{\sigma_{12} - \rho_1\rho_2\sigma_1\sigma_2}{\sigma_1\sqrt{1 - \rho_1^2} \;\cdot\; \sigma_2\sqrt{1 - \rho_2^2}} \tag{106}$$

satisfies $|\rho_{y|g}| < 1$, so the conditional posterior is a non-degenerate bivariate Gaussian. This ensures that the correlated HVI distribution machinery of Chapter 3 applies with parameters $(\boldsymbol{\mu}_{y|g}(t),\; \boldsymbol{\Sigma}_{y|g})$. In the degenerate boundary cases $\rho_i^2 \to 1$, the conditional variance in the i-th objective vanishes and the bivariate framework breaks down; these cases are excluded by the requirement that the joint covariance is positive definite.

Remark 4.16 (Induced dependence). The off-diagonal entry of $\boldsymbol{\Sigma}_{y|g}$ is:

$$\mathrm{Cov}(y_1, y_2 \mid g) = \sigma_{12} - \rho_1\rho_2\sigma_1\sigma_2. \tag{107}$$

Even when $\sigma_{12} = 0$ (marginally independent objectives), if both $\rho_1 \neq 0$ and $\rho_2 \neq 0$, then $\mathrm{Cov}(y_1, y_2 \mid g) = -\rho_1\rho_2\sigma_1\sigma_2 \neq 0$. That is, *conditioning on a common correlated variable induces dependence between the objectives.* This is the fundamental reason why the correlated-constraint case requires the bivariate HVI distribution machinery developed in Chapter 3, rather than the factored product-of-marginals approach of Wang et al. (2024).

4.3.2 CDF via Law of Total Probability

Theorem 4.17 (CDF of constrained HVI: correlated single constraint)**.** Under the joint model (102), the CDF of the constrained generalized HVI is:

$$F_{\Delta_C}(\delta) = \big(1 - \mathrm{PoF}(\boldsymbol{x})\big) \cdot \mathbf{1}_{\{\delta \geq 0\}} + \mathrm{PoF}(\boldsymbol{x}) \cdot F_{\Delta|g\leq 0}(\delta), \tag{108}$$

where $\mathrm{PoF}(\boldsymbol{x}) = \Phi(-\mu_c/\sigma_c)$ and $F_{\Delta|g\leq 0}(\delta)$ is the CDF of the generalized HVI computed under the conditional distribution of $\boldsymbol{y} \mid g \leq 0$.

Proof. By the law of total probability:

$$F_{\Delta_C}(\delta) = \Pr(\Delta_C \leq \delta \mid g \leq 0)\,\mathrm{PoF} + \Pr(\Delta_C \leq \delta \mid g > 0)\,(1 - \mathrm{PoF}). \tag{109}$$

The infeasible branch yields $\mathbf{1}_{\{\delta\geq 0\}}$ since $\Delta_C = 0$ when $g > 0$. For the feasible branch, $\Delta_C = \Delta(\boldsymbol{y})$, but the distribution of $\boldsymbol{y}$ given $g \leq 0$ differs from its marginal due to the correlation. Hence $\Pr(\Delta(\boldsymbol{y}) \leq \delta \mid g \leq 0) = F_{\Delta|g\leq 0}(\delta)$. □

Remark 4.18. When $\rho_1 = \rho_2 = 0$, the conditioning does not change the distribution of $\boldsymbol{y}$, so $F_{\Delta|g\leq 0} = F_\Delta$ and (108) reduces to the independent formula (97).

The structural form of (108) is identical to the independent case: a mixture of an infeasible point mass and a feasible continuous component. The difference lies entirely in the term $F_{\Delta|g\leq 0}(\delta)$, which now accounts for the modified objective distribution induced by conditioning on feasibility.

4.3.3 Computing $F_{\Delta|g\leq 0}$ via Integration Over the Constraint

The conditional HVI distribution is computed by integrating over the conditioning variable:

$$F_{\Delta|g\leq 0}(\delta) = \frac{1}{\mathrm{PoF}(\boldsymbol{x})} \int_{-\infty}^{0} F_{\Delta|g=t}(\delta)\ \varphi_{\mu_c,\sigma_c}(t)\,dt, \tag{110}$$

$$f_{\Delta|g\leq 0}(\delta) = \frac{1}{\mathrm{PoF}(\boldsymbol{x})} \int_{-\infty}^{0} f_{\Delta|g=t}(\delta)\ \varphi_{\mu_c,\sigma_c}(t)\,dt, \tag{111}$$

where $F_{\Delta|g=t}$ and $f_{\Delta|g=t}$ are the CDF and density of $\Delta(\boldsymbol{y})$ when $\boldsymbol{y}$ follows the conditional bivariate Gaussian from Lemma 4.14. The integrands are well-defined: since $f_{\Delta|g=t}(\delta)$ depends on t only through the affine conditional mean (104), and the bivariate Gaussian density is smooth in its mean parameter, the map $t \mapsto f_{\Delta|g=t}(\delta)$ is continuous and bounded on $(-\infty, 0]$.

The key step is computing $f_{\Delta|g=t}(\delta)$. Since conditioning on $g = t$ yields a bivariate Gaussian for (y_1, y_2) with conditional covariance $\boldsymbol{\Sigma}_{y|g}$ that generically has nonzero off-diagonal entries (Remark 4.16), the Wang et al. cell-level density, which factors as a product of independent univariate truncated Gaussians, cannot be applied. Instead, the correlated HVI distribution developed in Chapter 3 is required.

4.3.4 Cell-Level Density with Bivariate Truncated Normal

Proposition 4.19 (Cell-level density under correlated constraint)**.** For a non-dominated cell $C(i, j)$ with $i + j \leq n$, define the transformed variables $z_1 = y_1^{(n+1-j)} - y_1$ and $z_2 = y_2^{(i)} - y_2$, so that

$\Delta^+(\boldsymbol{y})|_{C(i,j)} = z_1 z_2 + \gamma^{(i,j)}$ (Section 2.4). Given $g = t$, the vector (z_1, z_2) follows a bivariate Gaussian:

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \Bigg| g = t \;\sim\; \mathcal{N}\big(\boldsymbol{\nu}(t),\; \boldsymbol{\Sigma}_{y|g}\big), \tag{112}$$

where:

$$\boldsymbol{\nu}(t) = \begin{pmatrix} y_1^{(n+1-j)} - \mu_1 - \rho_1 \dfrac{\sigma_1}{\sigma_c}(t - \mu_c) \\ y_2^{(i)} - \mu_2 - \rho_2 \dfrac{\sigma_2}{\sigma_c}(t - \mu_c) \end{pmatrix} =: \begin{pmatrix} \nu_1(t) \\ \nu_2(t) \end{pmatrix}, \tag{113}$$

and $\boldsymbol{\Sigma}_{y|g}$ is from (105). The covariance of (z_1, z_2) equals $\boldsymbol{\Sigma}_{y|g}$ because the transformation $\boldsymbol{y} \mapsto \boldsymbol{a} - \boldsymbol{y}$ has Jacobian $-\boldsymbol{I}$, and $(-\boldsymbol{I})\boldsymbol{\Sigma}_{y|g}(-\boldsymbol{I})^\top = \boldsymbol{\Sigma}_{y|g}$ (cf. Proposition 2.28).

Within cell $C(i,j)$, the variables z_1 and z_2 are restricted to the truncation rectangle $\mathcal{R} = [L_1, U_1] \times [L_2, U_2]$ from Section 2.5. The cell probability under the conditional distribution is:

$$P_C^{(i,j)}(t) := \Pr\big((z_1, z_2) \in \mathcal{R} \mid g = t\big) = \Phi_2\big(\boldsymbol{L},\, \boldsymbol{U};\; \boldsymbol{\nu}(t),\, \boldsymbol{\Sigma}_{y|g}\big), \tag{114}$$

where Φ_2 denotes the bivariate normal rectangle probability.

The density of $\Delta^+(\boldsymbol{y})|_{C(i,j)}$ given $g = t$, conditional on $\boldsymbol{y} \in C(i,j)$, is:

$$\mathrm{PDF}^{(i,j)}_{\Delta|C(i,j),\, g=t}(\delta) = \frac{1}{P_C^{(i,j)}(t)} \int_{\alpha(p)}^{\beta(p)} \varphi_2\Big(x,\, \frac{p}{x};\; \boldsymbol{\nu}(t),\, \boldsymbol{\Sigma}_{y|g}\Big)\, |x|^{-1}\, dx, \tag{115}$$

where $p = \delta - \gamma^{(i,j)}$ and the integration bounds $\alpha(p), \beta(p)$ enforce $x \in [L_1, U_1]$ and $p/x \in [L_2, U_2]$ simultaneously, following the case analysis of Section 2.6.

Proof. The transformation $\boldsymbol{y} \mapsto \boldsymbol{z} = \boldsymbol{a} - \boldsymbol{y}$ with $\boldsymbol{a} = (y_1^{(n+1-j)},\, y_2^{(i)})^\top$ is affine with Jacobian $-\boldsymbol{I}$. By Lemma 4.14, $\boldsymbol{y} \mid g = t$ is bivariate Gaussian, so $\boldsymbol{z} \mid g = t$ is also bivariate Gaussian with mean $\boldsymbol{a} - \boldsymbol{\mu}_{y|g}(t) = \boldsymbol{\nu}(t)$ and covariance $\boldsymbol{\Sigma}_{y|g}$.

Within cell $C(i,j)$, the HVI equals $z_1 z_2 + \gamma^{(i,j)}$ (Section 2.4). The density of the product $W = z_1 z_2$ under the truncated bivariate normal on $\mathcal{R}$ follows from the change-of-variables lemma of Section 2.6: the substitution $(z_1, z_2) \mapsto (z_1, w)$ with $w = z_1 z_2$ has Jacobian $|z_1|^{-1}$. For a fixed product value p, the level set $\{z_1 z_2 = p\} \cap \mathcal{R}$ is parameterized by $z_1 = x$, $z_2 = p/x$, yielding (115). □

Remark 4.20 (Comparison with the independent case). In the unconstrained setting with independent objectives ($\boldsymbol{\Sigma}_{y|g}$ diagonal), the bivariate density φ_2 factors as a product of univariate densities:

$$\varphi_2(z_1, z_2;\; \boldsymbol{\nu}, \boldsymbol{\Sigma}_{y|g}) = \varphi_{\nu_1, \tilde{\sigma}_1}(z_1) \cdot \varphi_{\nu_2, \tilde{\sigma}_2}(z_2), \tag{116}$$

where $\tilde{\sigma}_i^2 = \sigma_i^2(1 - \rho_i^2)$, and the cell probability factors as $P_C^{(i,j)}(t) = D_1^{-1} D_2^{-1}$. This recovers exactly the Wang et al. formula (2024, Eq. 4). The bivariate formulation is therefore a strict generalization.

Corollary 4.21 (Reduction when only one objective correlates with g)**.** If exactly one of the correlations is zero (say $\rho_2 = 0$) and the objectives are marginally independent ($\sigma_{12} = 0$), then:

$$\boldsymbol{\Sigma}_{y|g} = \begin{pmatrix} \sigma_1^2(1 - \rho_1^2) & 0 \\ 0 & \sigma_2^2 \end{pmatrix}. \tag{117}$$

The conditional covariance is diagonal, so the bivariate density factors and the Wang et al. product formula applies within each cell, with a shifted mean for y_1 only. This is the one special case where the correlated-constraint analysis does not require the full bivariate truncated normal machinery.

4.3.5 Dominated Cells Under Correlation

Proposition 4.22 (Dominated cell density under correlation)**.** For a dominated cell $C(i,j)$ with $i+j>n$, define the inverted variables:

$$z_1^- = -z_1 = y_1 - y_1^{(n+1-j)}, \qquad z_2^- = -z_2 = y_2 - y_2^{(i)}. \tag{118}$$

Then $(z_1^-, z_2^-)^\top \mid g = t \sim \mathcal{N}(-\boldsymbol{\nu}(t), \boldsymbol{\Sigma}_{y|g})$, with the same covariance $\boldsymbol{\Sigma}_{y|g}$ (since negating both variables preserves the covariance matrix). The cell bounds become $[L_k^-, U_k^-]$ with $L_k^- = -U_k$ and $U_k^- = -L_k$, and the negative HVI within the cell is:

$$\Delta^-(\boldsymbol{y})\big|_{C(i,j)} = -\big(z_1^- z_2^- + \gamma^{-(i,j)}\big), \tag{119}$$

where $\gamma^{-(i,j)} = -\Delta^-(\boldsymbol{l}^{(i,j)}) - (l_1^{(i,j)} - y_1^{(n+1-j)})(l_2^{(i,j)} - y_2^{(i)})$, matching the offset defined in (44). The conditional density satisfies:

$$\mathrm{PDF}^{(i,j)}_{\Delta^-|C(i,j),\,g=t}(\delta) = \mathrm{PDF}^{(i,j)}_{\Delta^+|C(i,j),\,g=t}(-\delta), \tag{120}$$

evaluated with the inverted mean $-\boldsymbol{\nu}(t)$, inverted bounds, and the same covariance $\boldsymbol{\Sigma}_{y|g}$ in (115).

Proof. The map $(z_1, z_2) \mapsto (z_1^-, z_2^-) = (-z_1, -z_2)$ has Jacobian $(-\boldsymbol{I})$, so $\mathrm{Cov}(z_1^-, z_2^-) = (-\boldsymbol{I})\boldsymbol{\Sigma}_{y|g}(-\boldsymbol{I})^\top = \boldsymbol{\Sigma}_{y|g}$, and the mean becomes $-\boldsymbol{\nu}(t)$. The cell bounds invert as stated. Since $z_1^- z_2^- = z_1 z_2$, the product $W^- = z_1^- z_2^-$ has the same distribution as $W = z_1 z_2$ with the substituted parameters. The negative HVI is $-W^- - \gamma^{-(i,j)}$, so its density at δ equals the density of W^- at $-\delta - \gamma^{-(i,j)}$, which matches the positive-HVI formula with the inverted quantities. □

4.3.6 Assembling the Full Conditional CDF

The full density of HVI given $g = t$ is assembled over all cells. Each cell-level density $\mathrm{PDF}^{(i,j)}_{\Delta|C(i,j),g=t}$ is conditional on $\boldsymbol{y} \in C(i,j)$, so the mixture uses cell probabilities as weights. The normalization factor $1/P_C^{(i,j)}(t)$ inside the conditional density cancels with the weight $P_C^{(i,j)}(t)$, yielding the unconditional contribution from each cell:

$$f_{\Delta|g=t}(\delta) = \sum_{i,j\in[0..n]} \int_{\alpha^{(i,j)}(p)}^{\beta^{(i,j)}(p)} \varphi_2\Big(x,\, \frac{p}{x};\, \boldsymbol{\nu}^{(i,j)}(t),\, \boldsymbol{\Sigma}_{y|g}\Big)\, |x|^{-1}\, dx, \tag{121}$$

where $p = \delta - \gamma^{(i,j)}$ for each cell, using the appropriate formula for non-dominated cells (Proposition 4.19) or dominated cells (Proposition 4.22), and only cells whose HVI range contains δ contribute.

The conditional CDF is obtained by numerical integration of the density:

$$F_{\Delta|g=t}(\delta) = \int_{-\infty}^{\delta} f_{\Delta|g=t}(\delta')\, d\delta'. \tag{122}$$

No closed-form double integral is presented because the product range and integration bounds require the full case analysis of the product density change-of-variables (Section 2.6), which depends on sign quadrants within each cell. The conditional CDF given $g \le 0$ is then obtained via (110), and the full constrained CDF via (108).

4.3.7 Numerical Computation

Evaluating the correlated density (111) requires a nested integration: an outer integral over $t \in (-\infty, 0]$ and, for each value of t, the inner cell-level integral from the bivariate normal product density.

Outer integral (over the truncated constraint). The outer integral is over $g \mid \mathcal{D} \sim \mathcal{N}(\mu_c, \sigma_c^2)$ restricted to $(-\infty, 0]$. We transform to a standard variable via $t = \mu_c + \sigma_c s$, where s ranges over $(-\infty, s^*]$ with $s^* = -\mu_c/\sigma_c$. To apply standard quadrature on a finite interval, we use the inverse-CDF transformation: let $u \in [0, 1]$ and define $s(u) = \Phi^{-1}(u \cdot \Phi(s^*))$, so that $t(u) = \mu_c + \sigma_c \cdot s(u)$. The integral becomes:

$$\int_{-\infty}^{0} f_{\Delta|g=t}(\delta)\, \varphi_{\mu_c,\sigma_c}(t)\, dt = \mathrm{PoF}(\boldsymbol{x}) \cdot \int_{0}^{1} f_{\Delta|g=t(u)}(\delta)\, du, \tag{123}$$

which is a standard integral on $[0, 1]$, amenable to Gauss-Legendre quadrature with K nodes:

$$f_{\Delta|g\leq 0}(\delta) \approx \sum_{k=1}^{K} w_k\, f_{\Delta|g=t(u_k)}(\delta), \tag{124}$$

where $\{u_k, w_k\}_{k=1}^{K}$ are the Gauss–Legendre nodes and weights on $[0, 1]$, and $t(u_k) = \mu_c + \sigma_c \cdot \Phi^{-1}(u_k \cdot \Phi(s^*))$.

Inner integral (cell-level bivariate product density). For each quadrature node t_k, we evaluate the bivariate cell density (115) using Gauss–Kronrod quadrature (21-point rule with up to 50 sub-intervals, following Wang et al.) for the integral over the hyperbola parameterization.

Bivariate rectangle probabilities. The cell probabilities $P_C^{(i,j)}(t)$ are bivariate normal rectangle probabilities. We compute these using the algorithm of [11], which reduces the bivariate CDF to a one-dimensional integral evaluated by Gauss–Legendre quadrature with guaranteed absolute error below 10^{-15}.

Complexity. The overall complexity per candidate point $\boldsymbol{x}$ is $O(K \cdot n_{\text{active}}^2 \cdot \tau^{-1})$, where K is the number of outer quadrature nodes, n_{active} is the number of active cells (after pruning cells with negligible probability mass, as in Wang et al.), and τ^{-1} is the cost of the inner Gauss-Kronrod rule. In practice, $K = 15$–20 nodes suffice for the outer integral, and each bivariate rectangle probability adds $O(1)$ per cell evaluation.

4.4 Extension to J Inequality Constraints

The single-constraint results are now extended to the general case of J inequality constraints $c_j(\boldsymbol{x}) \leq 0$ for $j \in [1..J]$. No new geometry is introduced in this section; the cell partition and HVI structure are unchanged from Sections 2.5 and 2.4. The extension is purely probabilistic, concerning the computation of the feasibility probability and the conditional objective distribution under multiple constraints.

4.4.1 Independent Constraints and Objectives

Assumption 4.23 (Full independence)**.** All objectives and constraints are mutually independent given $\mathcal{D}$:

$$y_1 \perp y_2 \perp g_1 \perp \cdots \perp g_J \mid \mathcal{D}. \tag{125}$$

Under this assumption, the feasibility indicator $I_{\text{feas}} = \prod_{j=1}^{J} \mathbf{1}_{\{g_j \le 0\}}$ is independent of $\boldsymbol{y}$, and the joint probability of feasibility factorizes:

$$\text{PoF}_{\text{joint}}(\boldsymbol{x}) = \Pr(g_1 \le 0, \ldots, g_J \le 0 \mid \mathcal{D}, \boldsymbol{x}) = \prod_{j=1}^{J} \text{PoF}_j(\boldsymbol{x}), \tag{126}$$

where $\text{PoF}_j(\boldsymbol{x}) = \Phi(-\mu_{c_j}/\sigma_{c_j})$.

Corollary 4.24 (CDF under J independent constraints)**.** Under Assumption 4.23, the CDF of Δ_{C} is:

$$F_{\Delta_{\text{C}}}(\delta) = \left(1 - \prod_{j=1}^{J} \text{PoF}_j(\boldsymbol{x})\right) \cdot \mathbf{1}_{\{\delta \ge 0\}} + \left(\prod_{j=1}^{J} \text{PoF}_j(\boldsymbol{x})\right) \cdot F_{\Delta}(\delta), \tag{127}$$

where $F_{\Delta}(\delta)$ is the unconstrained HVI CDF.

Proof. Under mutual independence, $I_{\text{feas}} = \prod_j \mathbf{1}_{\{g_j \le 0\}}$ is independent of $\boldsymbol{y}$, and $\Pr(I_{\text{feas}} = 1) = \prod_j \text{PoF}_j$. The proof of Theorem 4.9 applies with PoF replaced by $\prod_j \text{PoF}_j$. □

The corresponding density for $\delta \neq 0$ is:

$$f_{\Delta_{\text{C}}}(\delta) = \left(\prod_{j=1}^{J} \text{PoF}_j(\boldsymbol{x})\right) \cdot f_{\Delta}(\delta). \tag{128}$$

This result is a direct generalization of the single independent constraint case from Section 4.2: the unconstrained HVI distribution is simply rescaled by the product of individual feasibility probabilities. No integration over the constraint space is required.

4.4.2 Correlated Constraints and Objectives

For the general correlated case, define the joint posterior over objectives and all constraints:

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{g} \end{pmatrix} \Bigg| \mathcal{D}, \boldsymbol{x} \;\sim\; \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_g \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yg} \\ \boldsymbol{\Sigma}_{gy} & \boldsymbol{\Sigma}_{gg} \end{pmatrix} \right), \tag{129}$$

where $\boldsymbol{y} = (y_1, y_2)^\top \in \mathbb{R}^2$, $\boldsymbol{g} = (g_1, \ldots, g_J)^\top \in \mathbb{R}^J$, and $\boldsymbol{\Sigma}_{yg} \in \mathbb{R}^{2 \times J}$ captures all objective-constraint cross-covariances.

Theorem 4.25 (CDF under J correlated constraints)**.** The CDF of Δ_{C} under (129) is:

$$F_{\Delta_{\text{C}}}(\delta) = \big(1 - \text{PoF}_{\text{joint}}(\boldsymbol{x})\big) \cdot \mathbf{1}_{\{\delta \ge 0\}} + \text{PoF}_{\text{joint}}(\boldsymbol{x}) \cdot F_{\Delta \mid \boldsymbol{g} \preceq \mathbf{0}}(\delta), \tag{130}$$

where:

$$\text{PoF}_{\text{joint}}(\boldsymbol{x}) := \Pr(g_1 \le 0, \ldots, g_J \le 0 \mid \mathcal{D}, \boldsymbol{x}) \tag{131}$$

is the Gaussian orthant probability of the J-variate distribution $\boldsymbol{g} \mid \mathcal{D}, \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{gg})$, and $F_{\Delta \mid \boldsymbol{g} \preceq \mathbf{0}}(\delta)$ is the CDF of HVI under the conditional $\boldsymbol{y} \mid \boldsymbol{g} \preceq \mathbf{0}, \mathcal{D}, \boldsymbol{x}$.

Proof. Identical structure to Theorem 4.17, conditioning on $\{\boldsymbol{g} \preceq \mathbf{0}\}$. □

Conditional objective distribution. By multivariate Gaussian conditioning (Section 2.3), the distribution of $\boldsymbol{y}$ given $\boldsymbol{g} = \boldsymbol{t}$ is:

$$\boldsymbol{y} \mid \boldsymbol{g} = \boldsymbol{t}, \mathcal{D} \;\sim\; \mathcal{N}\big(\boldsymbol{\mu}_y(\boldsymbol{t}),\; \boldsymbol{\Sigma}_{y|g}\big), \tag{132}$$

where:

$$\boldsymbol{\mu}_y(\boldsymbol{t}) = \boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yg}\boldsymbol{\Sigma}_{gg}^{-1}(\boldsymbol{t} - \boldsymbol{\mu}_g), \tag{133}$$

$$\boldsymbol{\Sigma}_{y|g} = \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yg}\boldsymbol{\Sigma}_{gg}^{-1}\boldsymbol{\Sigma}_{gy}. \tag{134}$$

As in the single-constraint case, $\boldsymbol{\Sigma}_{y|g}$ is a 2×2 matrix that is generically not diagonal. The conditional mean $\boldsymbol{\mu}_y(\boldsymbol{t})$ is affine in $\boldsymbol{t}$, and $\boldsymbol{\Sigma}_{y|g}$ is independent of $\boldsymbol{t}$.

Remark 4.26 (Dependence of both factors on $\boldsymbol{\Sigma}_{yg}$). Both factors in the ε-CPoHVI expression (139) (Section 4.5) depend on the cross-covariance $\boldsymbol{\Sigma}_{yg}$: the joint feasibility probability $\mathrm{PoF}_{\text{joint}}$ depends on $\boldsymbol{\Sigma}_{gg}$ (which may itself be influenced by the multi-output GP structure that also generates $\boldsymbol{\Sigma}_{yg}$), and the conditional HVI CDF $F_{\Delta|\boldsymbol{g}\preceq\mathbf{0}}$ depends on $\boldsymbol{\Sigma}_{yg}$ through the conditional mean (133) and covariance (134). In the independent case, $\boldsymbol{\Sigma}_{yg} = \mathbf{0}$, so $F_{\Delta|\boldsymbol{g}\preceq\mathbf{0}} = F_\Delta$ and $\mathrm{PoF}_{\text{joint}} = \prod_j \mathrm{PoF}_j$, recovering the product formula of Corollary 4.24.

Cell-level density under J correlated constraints. The cell-level density of HVI given $\boldsymbol{g} = \boldsymbol{t}$ takes the same bivariate normal product form as Proposition 4.19, with the parameters replaced by conditional quantities from (133)-(134):

$$\mathrm{PDF}^{(i,j)}_{\Delta|C(i,j),\boldsymbol{g}=\boldsymbol{t}}(\delta) = \frac{1}{P_C^{(i,j)}(\boldsymbol{t})} \int_{\alpha(p)}^{\beta(p)} \varphi_2\Big(x,\, \frac{p}{x};\; \boldsymbol{\nu}^{(i,j)}(\boldsymbol{t}),\, \boldsymbol{\Sigma}_{y|g}\Big)\, |x|^{-1}\, dx, \tag{135}$$

where $\boldsymbol{\nu}^{(i,j)}(\boldsymbol{t}) = (y_1^{(n+1-j)},\, y_2^{(i)})^\top - \boldsymbol{\mu}_y(\boldsymbol{t})$ and $P_C^{(i,j)}(\boldsymbol{t}) = \Phi_2(\boldsymbol{L}^{(i,j)},\, \boldsymbol{U}^{(i,j)};\; \boldsymbol{\nu}^{(i,j)}(\boldsymbol{t}),\, \boldsymbol{\Sigma}_{y|g})$.

The unconditional density given $\boldsymbol{g} \preceq \mathbf{0}$ is:

$$f_{\Delta|\boldsymbol{g}\preceq\mathbf{0}}(\delta) = \frac{1}{\mathrm{PoF}_{\text{joint}}} \int_{\mathbb{R}_-^J} \left[\sum_{i,j} \int_{\alpha^{(i,j)}(p)}^{\beta^{(i,j)}(p)} \varphi_2\Big(x,\, \frac{p}{x};\; \boldsymbol{\nu}^{(i,j)}(\boldsymbol{t}),\, \boldsymbol{\Sigma}_{y|g}\Big)\, |x|^{-1}\, dx \right] \varphi_{\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{gg}}(\boldsymbol{t})\, d\boldsymbol{t}, \tag{136}$$

where $\mathbb{R}_-^J = (-\infty, 0]^J$ and $p = \delta - \gamma^{(i,j)}$ for each cell.

Computation of the joint probability of feasibility. When the constraints are correlated ($\boldsymbol{\Sigma}_{gg}$ non-diagonal), $\mathrm{PoF}_{\text{joint}}$ is the orthant probability of a J-variate Gaussian. For $J = 1$, this is the univariate CDF. For $J = 2$, it is computed via the bivariate normal CDF [11]. For general J, the randomized quasi-Monte Carlo algorithm of [17] is used, which provides stochastic error bounds; implementations are available in `scipy.stats.mvn.mvnun` (Python) and the `mvtnorm` package (R). Under independent constraints ($\boldsymbol{\Sigma}_{gg}$ diagonal, $\boldsymbol{\Sigma}_{yg} = \mathbf{0}$), it factorizes as $\prod_j \mathrm{PoF}_j$.

Numerical integration over the feasible constraint region. For the J-dimensional outer integral in (136), we generalize the inverse-CDF approach from Section 4.3.7:

1. **Transform via Cholesky.** Let $\boldsymbol{g} = \boldsymbol{\mu}_g + \boldsymbol{L}\boldsymbol{s}$, where $\boldsymbol{L}\boldsymbol{L}^\top = \boldsymbol{\Sigma}_{gg}$ and $\boldsymbol{s} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_J)$. The feasible region becomes $\{\boldsymbol{s} : \boldsymbol{\mu}_g + \boldsymbol{L}\boldsymbol{s} \preceq \mathbf{0}\}$.

2. **Quasi-Monte Carlo sampling.** For moderate J ($J \leq 5$, typical in constrained Bayesian optimization), we draw K_{QMC} samples from the truncated J-variate standard normal restricted to the feasible region, using the separation-of-variables method of [16]. Each sample $\boldsymbol{s}_k$ yields $\boldsymbol{t}_k = \boldsymbol{\mu}_g + \boldsymbol{L}\boldsymbol{s}_k$ and contributes one evaluation of the inner cell-level density.

3. **Independent constraints.** When $\boldsymbol{\Sigma}_{gg}$ is diagonal and $\boldsymbol{\Sigma}_{yg} = \boldsymbol{0}$, no integration over the constraint space is needed; the result is the product formula of Corollary 4.24.

Complexity. The total cost per candidate is $O(K_{\text{QMC}} \cdot n_{\text{active}}^2 \cdot \tau^{-1})$, where K_{QMC} is the number of QMC samples for the J-dimensional truncated integral. Each bivariate rectangle probability adds $O(1)$ per cell evaluation via the Drezner–Wesolowsky algorithm.

4.5 The Constrained ε-PoHVI Acquisition Function

Having derived the exact distribution of the constrained generalized HVI in Sections 4.2-4.4, the constrained acquisition function is now defined and establish its key properties.

4.5.1 Definition

Definition 4.27 (Constrained ε-Probability of Hypervolume Improvement)**.** Given the CDF of the constrained generalized HVI, we define:

$$\varepsilon\text{-CPoHVI}(\boldsymbol{x};\, \mathcal{P},\, \boldsymbol{r},\, \varepsilon) := 1 - F_{\Delta_C|\mathcal{D},\boldsymbol{x}}(\varepsilon) = \Pr\big(\Delta_C(\boldsymbol{y}, \boldsymbol{g}) > \varepsilon \;\big|\; \mathcal{D}, \boldsymbol{x}\big). \tag{137}$$

This computes the probability that the candidate point $\boldsymbol{x}$ is both feasible and achieves more than ε hypervolume improvement.

4.5.2 Explicit Forms

Independent case. Substituting the CDF from Corollary 4.24 into (137), for $\varepsilon \geq 0$:

$$\begin{aligned}
\varepsilon\text{-CPoHVI}(\boldsymbol{x}) &= 1 - \left(1 - \prod_{j=1}^{J} \text{PoF}_j\right) - \prod_{j=1}^{J} \text{PoF}_j \cdot F_\Delta(\varepsilon) \\
&= \prod_{j=1}^{J} \text{PoF}_j(\boldsymbol{x}) \cdot \big(1 - F_\Delta(\varepsilon)\big) \\
&= \prod_{j=1}^{J} \text{PoF}_j(\boldsymbol{x}) \cdot \varepsilon\text{-PoHVI}(\boldsymbol{x};\, \mathcal{P},\, \boldsymbol{r},\, \varepsilon).
\end{aligned} \tag{138}$$

Remark 4.28 (Recovery of the Gardner et al. product form). Equation (138) shows that under independence, ε-CPoHVI equals the unconstrained ε-PoHVI multiplied by the product of feasibility probabilities. This is precisely the [14] weighting scheme applied to ε-PoHVI. The independent case therefore provides a theoretical justification for the product approach within the HVI distribution framework: under the independence assumption, the product weighting is exact and not an ad hoc heuristic.

Correlated case. Substituting the CDF from Theorem 4.25 into (137), for $\varepsilon \geq 0$:

$$\varepsilon\text{-CPoHVI}(\boldsymbol{x}) = \text{PoF}_{\text{joint}}(\boldsymbol{x}) \cdot \big(1 - F_{\Delta|\boldsymbol{g}\preceq\mathbf{0}}(\varepsilon)\big). \tag{139}$$

Structurally, this still factorizes as a feasibility probability times a conditional improvement probability. However, the two factors are not independent: both depend on the cross-covariance $\boldsymbol{\Sigma}_{yg}$ (Remark 4.26), and the conditional HVI CDF $F_{\Delta|\boldsymbol{g}\preceq\mathbf{0}}$ is computed under a modified objective distribution that accounts for the correlation with constraints. The product weighting formula (138) is not valid in this case; it would use the wrong (unconditional) HVI distribution.

4.5.3 Properties

Proposition 4.29 (Boundary cases)**.** The following limiting cases hold:

1. *No constraints* ($J = 0$): ε-CPoHVI $\equiv$ ε-PoHVI, recovering [27].
2. *Deterministic feasibility* ($\text{PoF}_{\text{joint}} = 1$): ε-CPoHVI $= 1 - F_{\Delta}(\varepsilon) = \varepsilon$-PoHVI.
3. *Deterministic infeasibility* ($\text{PoF}_{\text{joint}} = 0$): ε-CPoHVI $= 0$ for all ε.
4. $\varepsilon = 0$: ε-CPoHVI$(\boldsymbol{x}) = \text{PoF}_{\text{joint}}(\boldsymbol{x}) \cdot \Pr(\Delta > 0 \mid \boldsymbol{g} \preceq \mathbf{0}, \mathcal{D}, \boldsymbol{x})$, the probability of feasible strict improvement.
5. $\varepsilon \to \infty$: ε-CPoHVI $\to 0$.
6. $\varepsilon < 0$ (generalized HVI): Accounts for feasible dominated points whose negative HVI exceeds the threshold, providing a nontrivial acquisition landscape even when most of the objective space is dominated.

Proof. Cases 1-3 follow by substituting the respective CDF expressions. Cases 4-5 follow from the monotonicity of $F_{\Delta_{\text{C}}}$. Case 6 follows from the continuity of the generalized HVI distribution over negative values (Section 4.1). □

Proposition 4.30 (Monotonicity in ε)**.** ε-CPoHVI$(\boldsymbol{x}; \varepsilon)$ is non-increasing in ε:

$$\varepsilon_1 \leq \varepsilon_2 \quad \Longrightarrow \quad \varepsilon\text{-CPoHVI}(\boldsymbol{x}; \varepsilon_1) \geq \varepsilon\text{-CPoHVI}(\boldsymbol{x}; \varepsilon_2). \tag{140}$$

Proof. $F_{\Delta_{\text{C}}}(\delta)$ is non-decreasing in δ (as a CDF), so $1 - F_{\Delta_{\text{C}}}(\varepsilon)$ is non-increasing in ε. □

4.5.4 Relationship to Other Acquisition Functions

Proposition 4.31 (Constrained EHVI)**.** The constrained expected hypervolume improvement is recoverable from the distribution:

$$\text{CEHVI}(\boldsymbol{x}) = \mathbb{E}[\Delta_{\text{C}}] = \int_0^{\infty} \big(1 - F_{\Delta_{\text{C}}}(\delta)\big)\, d\delta - \int_{-\infty}^{0} F_{\Delta_{\text{C}}}(\delta)\, d\delta. \tag{141}$$

Under independence: $\text{CEHVI}(\boldsymbol{x}) = \prod_j \text{PoF}_j(\boldsymbol{x}) \cdot \text{EHVI}(\boldsymbol{x})$.

Proof. The first equality is the definition of expectation. The integral identity follows from the standard tail-sum formula for expectations, extended to random variables with negative support via the generalized HVI. Under independence, $F_{\Delta_C}(\delta) = (1 - \prod_j \mathrm{PoF}_j) \cdot \mathbf{1}_{\{\delta \geq 0\}} + \prod_j \mathrm{PoF}_j \cdot F_{\Delta}(\delta)$, so the point mass at zero drops out of the integral and the continuous part factors. □

Remark 4.32 (Relationship to ε-PoI). The ε-PoI acquisition function measures $\Pr(\boldsymbol{y} + \varepsilon \mathbf{1}_2 \in \mathrm{ndom}(\mathcal{P}))$, a binary indicator of proximity to the attainment boundary. In contrast, ε-CPoHVI uses the exact HVI distribution, providing a quantile-based measure of improvement quality under constraints. This carries two advantages: first, the HVI distribution captures how much improvement a point contributes, not merely whether it crosses a threshold; second, the constraint integration allows for joint reasoning about feasibility and improvement rather than treating them as separate factors.

5 Interpretation of Theoretical Results

The preceding two chapters broadened the Wang et al. [27] HVI distribution in two separate modeling directions: incorporating correlated objective posteriors (Chapter 3) and handling black-box inequality constraints (Chapter 4). Although the two extensions arose from different motivations, they share a single technical core: in both, the bivariate Gaussian posterior at a candidate point has a non-trivial off-diagonal covariance, and the cell integrals that build the HVI distribution must be evaluated under a dependent rather than a separable joint law. This chapter steps back from the derivations to translate those mechanics into modeling consequences: how does the correlated cell machinery change which candidates ε-PoHVI prefers? When does feasibility conditioning matter beyond a simple feasibility weight? And when, in practice, does the extra machinery actually pay off?

5.1 What Correlation Changes

Wang et al. [27] derive the HVI distribution under independent objective posteriors. Removing that assumption produces qualitative, not merely quantitative, changes to the mathematical structure of the result, and those structural changes have direct consequences for acquisition behavior.

Cell-level mass redistributes with ρ. Under independence, the probability that the posterior falls in a given cell factors into a product of two univariate Gaussian CDFs (Remark 3.7). When $\rho \neq 0$, this product form breaks down: the cell probability becomes a bivariate normal rectangle probability (Theorem 3.6) computed via four evaluations of $\Phi_2(\cdot, \cdot; \rho)$. In standardized objective coordinates, positive ρ concentrates posterior mass along same-sign deviations from the mean and negative ρ favors opposite-sign deviations; mapped onto the cell partition, this shifts which cells carry most of the mixture weight and therefore changes the shape of the overall HVI distribution. Concretely, for a candidate whose mean lies just below the attainment surface, positive ρ amplifies the probability of *simultaneous* improvement in both objectives and inflates the upper tail of the HVI distribution; negative ρ suppresses it.

The acquisition-level consequence is that ε-PoHVI evaluates a tail probability $\Pr(\Delta > \varepsilon)$, not a mean. Two candidates with identical *marginal* posteriors can therefore receive different ε-PoHVI values once their posterior correlations differ, because the tail mass is supplied by different cells in each case. Cell-level redistribution that leaves the posterior mean and marginal variances unchanged can still flip the candidate ranking.

Correlation destroys separability inside each cell. Within a cell, the HVI is a product of two transformed Gaussian variables plus a cell-specific constant. Under independence, the exponent in the product-density integrand separates into a function of x alone plus a function of p/x alone (Proposition 2.33); the integrand is therefore the product of two univariate Gaussian densities and the Wang et al. integral is one-dimensional and modular in the two objectives. The substitution $z_2 = p/z_1$ already produces a Laurent-polynomial dependence on x: even under independence the marginal quadratic term $(z_2 - \nu_2)^2/\sigma_2^2$ contributes an x^{-2} piece. Correlation does not create that structure, but it changes the coefficients. When $\rho \neq 0$ a cross-term $-2\rho(x - \nu_1)(p/x - \nu_2)/(\sigma_1\sigma_2)$ enters the quadratic form, modifies the x, x^{-1}, and x^{-2} coefficients of the resulting Laurent-

polynomial exponent $a_2x^2 + a_1x + a_0 + b_1x^{-1} + b_2x^{-2}$ (Proposition 2.33), and more importantly prevents the exponent from separating into a function of x alone plus a function of p/x alone.

Geometrically, the integration runs along the rectangular hyperbola $z_2 = p/z_1$ inside the cell rectangle. Under independence, the bivariate density is a product, so the integrand factors into "density along z_1 times density along z_2." Under correlation the density's principal axes are rotated relative to the hyperbola, and different points on the hyperbola sit at different Mahalanobis distances from the mean in a way that cannot be untangled into separate single-axis contributions. Correlation eliminates the modular-in-objectives structure that made the independent derivation clean.

Setting $\rho = 0$ recovers Wang et al. [27] exactly (Theorem 3.20): the cell probability product factors because $\Phi_2(a, b; 0) = \Phi(a)\Phi(b)$, the cross-terms vanish from the Laurent coefficients, and the bivariate density factors into a product of marginals. At the opposite end of the range, the Laurent coefficients carry a $1/(1-\rho^2)$ factor that grows without bound as $|\rho| \to 1$, and (in the constrained setting of Section 5.2) the conditional covariance becomes near-singular as $|\rho_i| \to 1$. The numerical verification in Section 6.3 is therefore run only up to $\rho = 0.9$, and no claim is made about behavior in the degenerate limit.

The correlated formulation preserves the cell-partition structure: the number of cell contributions still grows as $\mathcal{O}((n+1)^2)$ in the size of the approximation set. What changes is the per-cell cost. Φ_2 replaces a product of Φ values, and the within-cell quadrature evaluates the full bivariate Gaussian density rather than a product of marginals. In the implementation used here this translates into a roughly 2 to 3× overhead per HVI evaluation; the practical timing impact is reported in Section 8.3.

5.2 What Constraints Change

Constraints do two conceptually different things to the HVI distribution: feasibility uncertainty adds a discrete mass at zero, while objective-constraint correlation reshapes the continuous feasible component. Chapter 4 works through these in the four cases (independent / correlated × single / multiple constraints); the modeling consequences follow.

In the unconstrained setting, generalized HVI is a continuous random variable whose distribution is a mixture over cells (Theorem 3.18). Once feasibility enters, the constrained HVI becomes a mixture of two qualitatively different components: a point mass at zero with weight $1 - \mathrm{PoF}_{\text{joint}}$ (the candidate is infeasible) and the conditional HVI distribution given feasibility,

$$F_{\Delta_C}(\delta) = (1 - \mathrm{PoF}_{\text{joint}}) \cdot \mathbf{1}_{\{\delta \geq 0\}} + \mathrm{PoF}_{\text{joint}} \cdot F_{\Delta|\boldsymbol{g} \preceq \mathbf{0}}(\delta).$$

The point mass is the discrete consequence of feasibility uncertainty; it is present even under independence and is what an ε-CPoHVI threshold acquisition must account for when it asks "what is the probability of *feasible* improvement exceeding ε?"

When the constraints are independent of the objectives, conditioning on feasibility leaves the objective distribution unchanged: $F_{\Delta|\boldsymbol{g} \preceq \mathbf{0}} = F_{\Delta}$. The acquisition then becomes the rescaled product

$$\varepsilon\text{-CPoHVI}(\boldsymbol{x}) = \mathrm{PoF}_{\text{joint}}(\boldsymbol{x}) \cdot \varepsilon\text{-PoHVI}(\boldsymbol{x}),$$

where $\mathrm{PoF}_{\text{joint}}$ is the joint feasibility probability. If the constraints are also mutually independent, this joint probability factors as $\prod_{j=1}^{J} \mathrm{PoF}_j(\boldsymbol{x})$, recovering the Gardner et al. [14] feasibility weighting (Corollary 4.24). The contribution of Chapter 4 is to show that this familiar product form is

not an ad-hoc heuristic but an *exact* consequence of the HVI-distribution framework, under both objective-constraint and constraint-constraint independence. Outside that regime it ceases to be exact.

When objectives and constraints are modeled jointly, most naturally when a multi-output GP encodes a physical relationship between them into the surrogate, conditioning on feasibility changes the objective law. For each fixed constraint value $g = t$, Gaussian conditioning gives a bivariate Gaussian (Lemma 4.14)

$$\boldsymbol{y} \mid g = t \;\sim\; \mathcal{N}\big(\boldsymbol{\mu}_{y|g}(t),\; \boldsymbol{\Sigma}_{y|g}\big),$$

where the conditional covariance contains the off-diagonal term $\sigma_{12} - \rho_1\rho_2\sigma_1\sigma_2$. Here ρ_i denotes the objective-constraint correlation, not the objective-objective correlation ρ used in Section 5.1. This term is generically nonzero even when the unconditional objectives are marginally independent (Remark 4.16): two variables that are marginally independent can become dependent once one conditions on a shared selection variable. This is analogous to a Berkson or selection effect [5]: here the selection event is feasibility, $g \leq 0$.

The feasible-event law $\boldsymbol{y} \mid g \leq 0$, however, is not a single Gaussian with this covariance; it is obtained by integrating the fixed-t conditional Gaussian laws over the truncated distribution of g. This is why product weighting is no longer exact under objective-constraint correlation: the product formula evaluates the tail probability under the *unconditional* objective law, whereas ε-CPoHVI evaluates the tail probability under the *feasible-event* law. The latter requires the correlated HVI machinery of Chapter 3 *inside* the outer feasibility-conditioning integral, with parameters $\boldsymbol{\mu}_{y|g}(t)$ and $\boldsymbol{\Sigma}_{y|g}$ at each fixed t. The discrepancy between product weighting and ε-CPoHVI is driven by how strongly feasibility conditioning reshapes the objective law and is expected to be most consequential where feasibility and improvement are both uncertain, typically near a constraint boundary that sits close to the Pareto-relevant region.

Returning to the materials engineer of the introduction: compliance and density are physically coupled through the underlying microstructure, and peak stress under load is a constraint that depends on the same microstructural variables. Then σ_{12}, ρ_1, and ρ_2 are all generically nonzero, and the conditioning correction $\sigma_{12} - \rho_1\rho_2\sigma_1\sigma_2$ is the off-diagonal entry of the fixed-$g{=}t$ conditional covariance used inside the feasibility-conditioning integral. Product weighting ignores that correction altogether; ε-CPoHVI does not.

5.3 Benefits of the Theory

The two contributions of this thesis, the correlated HVI distribution (Chapter 3) and the constrained HVI distribution (Chapter 4), together close the two limitations identified by Wang et al. [27, Section 6]. The focus in this section is on when each one matters in practice and on what the theory does not yet promise.

The correlated formulation is most useful when the surrogate learns a non-negligible posterior cross-covariance in regions that matter for the ε-tail of the HVI distribution. If the fitted $\rho(\boldsymbol{x})$ is near zero, or if non-zero $\rho(\boldsymbol{x})$ only affects cells with negligible probability or negligible HVI range, the correlated acquisition collapses effectively to the independent one. Conversely, when posterior dependence shifts probability mass into high-improvement cells, ε-PoHVI can rank candidates differently from the independent baseline even when the marginal means and variances are similar. This prediction is tested empirically in Section 8, where the benchmark is set up to separate

problems designed to expose cross-objective coupling from problems where the fitted ICM posterior learns only weak posterior correlation.

The mirror diagnostic for ε-CPoHVI is whether feasibility conditioning meaningfully reshapes the objective law at the candidates the acquisition optimizer visits. The discrepancy between ε-CPoHVI and Gardner-style product weighting is largest where the constraint boundary lies close to the Pareto-relevant region and feasibility is genuinely uncertain. In problems where the constraint is essentially inactive over the region the optimizer explores, the conditional law is close to the unconditional one and the two methods agree closely. Because this thesis derives and verifies ε-CPoHVI but does not benchmark it end-to-end (Section 6.3.3), the prediction is left as a hypothesis for future constrained BO experiments.

These results are distributional, not yet empirical. They show how the correct posterior HVI law changes under correlation and under feasibility conditioning. They do not by themselves imply that corr-ε-PoHVI or ε-CPoHVI will outperform their respective baselines on every problem; that depends on whether the fitted surrogate learns meaningful cross-output covariance and on whether that covariance affects the ε-tail of the HVI distribution at the candidates the acquisition optimizer actually visits. This distinction is what motivates the experimental chapters that follow.

The remaining question is whether the analytical CDFs derived in Chapters 3 and 4 are numerically stable enough to live inside acquisition functions, and whether exposing the full HVI distribution actually changes BO behavior on realistic problems. The next three chapters take up that question in sequence: verifying the distributions against Monte Carlo, selecting a multi-output surrogate that supplies meaningful posterior correlation, and benchmarking the resulting acquisition against the Wang et al. baseline on a suite of 13 bi-objective problems.

6 Acquisition Functions and Numerical Verification

The previous chapters derive the probability distribution of the hypervolume improvement under correlated Gaussian-process posteriors and under inequality constraints. This chapter turns those derivations into executable acquisition functions, introduces the test problems used by both verification and the BO benchmark, and verifies numerically that the analytical CDFs agree with Monte Carlo simulation to within the expected sampling error. This verification substantially reduces the likelihood that the BO results in later chapters are driven by errors in the HVI distribution computation.

6.1 Implementation

6.1.1 Software framework

The implementation extends the open-source MOBO framework of Wang et al. [27][1], which provides the cell-partition machinery, the independent HVI distribution, and the CMA-ES-based acquisition optimization loop. All new code (correlated HVI distribution, constrained HVI distribution, multi-output GP surrogates, and experiment automation) is our contribution.[2] The ε-PoHVI baseline used in the BO comparison of Chapter 8 is inherited unmodified from the authors' released repository. This is a deliberate choice: the correlated method is an upgrade path that should reduce to the baseline when its modeling assumptions are stripped, and inheriting the baseline verbatim reduces the risk that an implementation discrepancy is silently driving the comparison.

There is one important implementation detail: the released code uses a Matérn $\frac{3}{2}$ kernel (consistent with their `args.yml`), differing from the Matérn $\frac{5}{2}$ reported in the paper text. The released-code setting is used throughout so that the Chapter 8 empirical comparison is against the released-code ε-PoHVI baseline rather than a re-implementation that mixes the two.

At each iteration the framework normalizes the design variables to the unit hypercube (when bounds are available); fits a surrogate model to the observations; evaluates the acquisition function at candidate points using the full posterior covariance when available; maximizes the acquisition over the design space using CMA-ES; and evaluates the selected point on the true objective functions. The acquisition optimizer settings used in the BO experiments are documented in Section 8.1; this chapter does not exercise them.

6.1.2 Acquisition functions

Three acquisition functions appear in this thesis, all built on the ε-PoHVI tail-probability framework of Wang et al. [27]:

- **ε-PoHVI (indep)**: the independent HVI distribution paired with an independent-GP surrogate. At each candidate $\boldsymbol{x}$, the acquisition value is $\Pr[\Delta(\boldsymbol{y}) > \varepsilon_t]$, where the HVI distribution is computed from the two independent marginals (μ_1, σ_1^2) and (μ_2, σ_2^2).

- **corr-ε-PoHVI**: the correlated HVI distribution (Chapter 3) paired with a multi-output GP surrogate. At each candidate $\boldsymbol{x}$, the full bivariate Gaussian posterior $\boldsymbol{\Sigma}(\boldsymbol{x})$ enters the

[1] Available at https://github.com/wangronin/HVI-distribution.
[2] Available at https://github.com/rdorland/thesis.

analytical HVI CDF directly, exploiting the cross-objective coupling that independent GPs cannot represent. Theorem 3.20 proves analytically that setting $\rho \equiv 0$ in the correlated formula recovers the independent baseline; Section 6.3 confirms that the implementation of the correlated formula at $\rho = 0$ matches the implementation of the independent baseline to numerical precision (an implementation-correctness check, not a separate analytical theorem).

- **ε-CPoHVI**: the constrained HVI distribution (Chapter 4). This acquisition is verified numerically below but is not benchmarked in a full BO loop in this thesis; the reasoning and the natural future-work targets are documented in the closing paragraph of Section 6.3.

The ε schedule (a shared exponential decay) and the CMA-ES optimizer settings used to maximize these acquisitions are defined in Section 8.1, since they are properties of the BO loop rather than of the acquisition formulas themselves.

6.1.3 Numerical methods

The correlated HVI distribution introduces three numerical primitives:

1. **Bivariate normal CDF** $\Phi_2(a, b; \rho)$ evaluated by 20-point Gauss-Legendre quadrature using the Drezner-Wesolowsky algorithm [11]; this routine is expected to be highly accurate for the rectangle probabilities encountered here, and is checked indirectly through the Monte Carlo and deterministic sanity tests of Section 6.3.2.

2. **Within-cell product density** via `scipy.integrate.quad` with a Numba-compiled integrand, reusing the one-dimensional product-density parameterization of Wang et al. [27] now with the correlated Laurent-polynomial coefficients of Proposition 2.33.

3. **Conditional CDF** evaluated by direct integration of the product-density formula over the active cell range, avoiding numerical differentiation of the cell-level CDF and the associated boundary terms.

6.2 Test environments

The empirical chapters use two families of bi-objective test problems. They are introduced here because both the fitted-posterior verification in Section 6.3 and the BO benchmark in Chapter 8 rely on them. Reference points used inside the BO loop and for plotting are problem-specific properties of the BO experiment and live in Section 8.1.3; this chapter does not need them.

ZDT family [31]. ZDT1, ZDT2, ZDT3, ZDT4, and ZDT6 are bi-objective benchmarks with analytically known Pareto fronts of varying geometry: convex (ZDT1), concave (ZDT2), discontinuous (ZDT3), multimodal (ZDT4), and non-uniform with biased density (ZDT6). All use the standard form $f_1 = x_1$, $f_2 = g(\boldsymbol{x})\, h(f_1, g)$ where g is a sum or modulation over $x_{2:d}$. In random-design samples these objectives can show substantial raw output correlation, but this correlation mainly reflects the deterministic functional structure of the benchmark and does not necessarily imply useful posterior cross-covariance at a fixed candidate point. The thesis includes ZDT problems as well-understood baselines, not as the primary benchmark for the correlated method.

WOSGZ family [28]. WOSGZ1-8 are the distance-to-attainable-boundary problems of Wang, Ong, Sun, Gupta & Zhang. Each WOSGZ variant is parameterized by (A, B, C, D, E) controlling the amplitude and shape of the bias terms in a common decomposition

$$f_k(\boldsymbol{x}) = h_k(x_1) + \frac{A \sin^D\!\left(\frac{\pi}{2}|2x_1 - 1|\right) + 1}{|J_k|} \sum_{j \in J_k} \left|x_j - \delta_j(\boldsymbol{x})\right|^C, \tag{142}$$

where $h_1 = 1 - x_1$, $h_2 = x_1$ (or their squares for WOSGZ8), and δ_j is a trigonometric shift depending on x_1 and E; the index sets J_1, J_2 partition $\{2, \ldots, n_{\text{var}}\}$. The problems are designed to stress MOBO methods: their objectives carry genuine inter-objective coupling through the shared shape term, which is precisely the regime in which the correlated acquisition is expected to help. The implementation (`problems/wosgz.py`) was independently cross-checked against a second implementation by Santiago Garcia Mayayo; both paper-faithful implementations agree to machine precision on the test cases listed in `tests/test_wosgz.py`.

6.3 Numerical accuracy verification

6.3.1 Setup

The analytical HVI CDFs (independent, correlated, and constrained) are compared against Monte Carlo simulation across a grid of GP posterior parameters. For each configuration, $N = 10^5$ samples are drawn from the joint posterior, the empirical CDF $\widehat{F}_N$ is evaluated on a fixed grid $\mathcal{G}$ in δ (described in the reproducibility paragraph below), and the grid-maximum absolute deviation $\max_{\delta \in \mathcal{G}} |\widehat{F}_N(\delta) - F(\delta)|$ is recorded. This statistic underestimates the true Kolmogorov supremum by an amount governed by the grid spacing; the quantile-spaced grid with cell-range endpoints and the explicit $\delta = 0^{\pm}$ samples described below is dense enough that the underestimation is small relative to the empirical-CDF sampling scale, but the metric is reported and discussed below as a *grid-maximum* CDF error rather than as a true supremum.

Synthetic-grid parameter values. Table 1 lists the exact parameter values used by the synthetic verification cases. All synthetic cases use reference point $\boldsymbol{r} = (1.2, 1.2)$. Constraint configurations `easy_feasible` / `borderline` / `tight` achieve PoF $\approx 0.98/0.50/0.05$ respectively under $\Sigma_{yg} = 0$. Configurations with non-positive-definite joint covariance are rejected during construction by checking $\lambda_{\min}(\boldsymbol{\Sigma}) > 10^{-10}$; the per-case totals in Table 2 reflect the configurations that survive this check.

Table 1: Synthetic-grid parameter values used by Cases 1 to 7 and 9. Pareto-front geometries are taken from Wang et al. [27]; feasibility configurations are chosen to span easy/borderline/tight regimes.

Symbol	Values used
Pareto-front geometry	`simple` ($\{(0.2, 0.8), (0.8, 0.2)\}$), `moderate` (5-point), `sparse` ($\{(0.3, 0.6)\}$, single point)
Mean position $\boldsymbol{\mu}$	`near_front` $(0.4, 0.6)$, `well_inside` $(0.15, 0.15)$, `outside` $(0.9, 0.9)$, `asymmetric` $(0.2, 0.8)$
Marginal SD σ	$\{0.05, 0.20, 0.50\}$
ρ (Case 2)	$\{-0.9, -0.6, -0.3, +0.3, +0.6, +0.9\}$
Constraint regimes	$(\mu_g, \sigma_g) \in \{(-1.0, 0.5), (0.0, 0.5), (0.5, 0.3)\}$
Reference point	$\boldsymbol{r} = (1.2, 1.2)$
PD rejection rule	$\lambda_{\min}(\boldsymbol{\Sigma}_{\text{joint}}) > 10^{-10}$

Configurations tested. Nine cases are exercised:

- **Case 1: Unconstrained, independent** ($\rho = 0$): 3 Pareto-front geometries $\times$ 4 mean positions $\times$ 3 standard-deviation values = 36 configurations.
- **Case 2: Unconstrained, correlated** ($\rho \in \{-0.9, -0.6, -0.3, +0.3, +0.6, +0.9\}$): same base grid as Case 1, repeated for each of 6 correlation levels = 216 configurations. The grid is symmetric in sign; the boundary values $\rho = \pm 0.9$ are included as a stress check on the bivariate-CDF machinery.
- **Case 3: Constrained, objective-constraint independent** (i.e. $\Sigma_{yg} = 0$, three feasibility regimes with PoF $\approx 0.05, 0.50, 0.98$): 6 configurations.
- **Case 4: Constrained, with correlated objectives and independent constraint** (i.e. $\rho_{12} \in \{-0.6, +0.6\}$, $\Sigma_{yg} = 0$, three feasibility regimes): 12 configurations. This is the special case where the constraint is independent of the objectives.
- **Case 5: Constrained with nonzero objective-constraint correlation** ($\Sigma_{yg} \neq 0$, $J = 1$): a grid of $\rho_{yy} \in \{0, 0.3\}$ crossed with $(\rho_{y_1 g}, \rho_{y_2 g}) \in \{(0.3, 0), (0.3, 0.3), (-0.3, 0.3), (0.5, -0.3)\}$ at three feasibility regimes (PoF ≈ 0.05, ≈ 0.5 and ≈ 0.98), giving 24 configurations. The tight regime (PoF ≈ 0.05) is the one where the infeasible atom dominates and numerical issues are most likely; including it stresses the $\Sigma_{yg} \neq 0$ code path in the most challenging feasibility geometry. This case exercises the full objective-constraint conditioning path derived in Section 4.3. The analytical CDF is computed by Gauss-Legendre quadrature over the truncated constraint distribution, with each quadrature node feeding the conditional Gaussian parameters $\left(\boldsymbol{\mu}_{y|g}(t), \boldsymbol{\Sigma}_{y|g}\right)$ into a fresh correlated-HVI evaluation.
- **Case 6: Multiple independent constraints,** $J = 2$: 4 configurations exercising the product factorization $\text{PoF}_{\text{joint}} = \prod_{j=1}^{J} \text{PoF}_j$ of Corollary 4.24, crossed with both $\rho_{yy} = 0$ and $\rho_{yy} = 0.6$ on the objective side.

- **Case 7: Multiple correlated constraints,** $J = 2$: 4 configurations exercising the joint feasibility probability via the bivariate Gaussian orthant probability (with $\rho_{g_1 g_2} \in \{-0.5, +0.5\}$), crossed with both $\rho_{yy} \in \{0, 0.6\}$. Uses the `scipy.stats.multivariate_normal.cdf` orthant evaluation with a quasi-Monte Carlo fallback if needed. The constraint side is correlated, but $\Sigma_{yg} = 0$: the multivariate objective-constraint conditioning path is exercised separately by Case 9.

- **Case 8: Realistic ICM posteriors**: 200 configurations extracted from ICM models fitted on real training data, 20 candidate-point posteriors each from ZDT1, ZDT2, WOSGZ1-WOSGZ8. For each problem, an ICM model is fitted on $n_{\text{train}} = 30$ Latin-Hypercube training samples (Matérn 3/2 base kernel, Adam 1000 iterations, 3 random restarts). Candidate points are then selected from a separate 200-point Latin Hypercube pool by a deterministic stratified rule: the pool is sorted by $|\rho(\boldsymbol{x})|$ (the absolute posterior cross-correlation), and at each of the $\{0, 25, 50, 75, 100\}$th percentiles the four nearest pool indices are retained, giving $5 \times 4 = 20$ candidates per problem. The cell partition for each candidate's HVI distribution is built from the non-dominated subset of the (normalized) training observations of that problem, with reference point $\boldsymbol{r} = (1.2, 1.2)$ in the normalized objective space matching the synthetic-grid configurations. This case tests whether the analytical CDFs work on the kind of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ pairs the BO loop actually produces.

- **Case 9:** $J = 2$ **constraints with full objective-constraint correlation** ($\Sigma_{yg} \neq 0$, $\Sigma_{gg} \neq 0$, $J = 2$): a small grid varying $\rho_{g_1 g_2} \in \{-0.3, +0.3\}$, $\rho_{yy} \in \{0, 0.6\}$, and three Σ_{yg} correlation patterns at two feasibility regimes (PoF ≈ 0.5 and ≈ 0.05); 24 configurations remain after the positive-definiteness check of Table 1 (no proposed combination in this grid was rejected). This case exercises the multivariate conditioning path of Section 4.4.2: the analytical CDF is computed by tensor-product Gauss-Legendre quadrature over the bivariate truncated (g_1, g_2) joint, with conditional Gaussian parameters $(\boldsymbol{\mu}_{y|g}, \boldsymbol{\Sigma}_{y|g})$ at each node feeding a fresh correlated-HVI evaluation.

The grid totals **526 configurations**: $36 + 216$ unconstrained $+ 6 + 12 + 24 + 4 + 4 + 24$ constrained $= 326$ synthetic plus 200 fitted-posterior. Per-case counts appear in Table 2.

Acceptance threshold and Monte Carlo calibration. A configuration passes if the grid-max CDF error stays below 0.01. For $N = 10^5$ Monte Carlo samples, the pointwise standard error of the empirical CDF is at most $1/(2\sqrt{N}) \approx 0.0016$, and the Dvoretzky-Kiefer-Wolfowitz 95% uniform band on the empirical CDF is $\sqrt{\log(2/0.05)/(2N)} \approx 0.0043$. The 0.01 threshold is therefore conservative relative to Monte Carlo noise, with roughly 2.3× headroom over the per-configuration DKW band. Any family-wise statement across the full grid (with K configurations) uses the Bonferroni-adjusted DKW band $\sqrt{\log(2K/0.05)/(2N)}$, which for the present $K = 526$ grid is ≈ 0.0071, still well inside 0.01.

Reproducibility settings. $N = 10^5$ Monte Carlo samples per configuration; the CDF is evaluated on a quantile-spaced grid of ∼50 points spanning the support of the analytical CDF. For the constrained CDFs the grid explicitly includes $\delta = 0^-, 0, 0^+$ so the atom at $\delta = 0$ is sampled cleanly on both sides, and for unconstrained CDFs it includes the cell-range endpoints; the remaining points are quantile-spaced over the body of the distribution. Each configuration uses a deterministic

seed derived from a stable hash of its (case, PF, $\mu, \sigma, \rho, \ldots$) tuple, so configurations have independent Monte Carlo streams that remain reproducible without sharing state. Numerical primitives use the following tolerances: the within-cell product-density quadrature is `scipy.integrate.quad` with `epsabs` $= 10^{-6}$, `epsrel` $= 10^{-6}$, `limit` $= 50$; the bivariate normal CDF uses 20-point Gauss-Legendre with the Drezner-Wesolowsky algorithm; the $J = 1$ outer integral over the truncated constraint distribution uses 40-node Gauss-Legendre over the standardized interval $[-8, z_{\max}]$; the $J = 2$ outer integral uses tensor-product Gauss-Legendre with 15 nodes per axis over the rectangle $[-8, z_{1,\max}] \times [-8, z_{2,\max}]$. The $J = 2$ correlated-constraint case (Case 7) uses `scipy.stats.multivariate_normal.cdf` for the bivariate orthant probability; a deterministic Sobol QMC fallback with $M = 4096$ samples is wired in for robustness, but did not need to trigger on the present grid.

6.3.2 Results

Table 2: Numerical verification of analytical HVI CDFs against Monte Carlo simulation ($N = 10^5$ samples per configuration). For each configuration the empirical-CDF deviation $\max_{\delta \in \mathcal{G}} |\widehat{F}_N(\delta) - F(\delta)|$ is recorded over a quantile-spaced evaluation grid $\mathcal{G}$ of $\sim$50 points augmented with cell-range endpoints and $\delta = 0^{\pm}$. "Mean grid-max err" is the mean of this quantity over configurations; "Max grid-max err" is its maximum. Acceptance threshold is 0.01, comfortably above the per-configuration DKW 95% band ≈ 0.0043; all 526 configurations pass.

Case	Configs	Mean grid-max err	Max grid-max err
Unconstrained, indep	36	0.0019	0.0050
Unconstrained, corr	216	0.0016	0.0059
Constrained, indep	6	0.0013	0.0026
Constrained, corr objs ($\Sigma_{yg} = 0$)	12	0.0015	0.0034
Constrained, corr obj-constr ($\Sigma_{yg} \neq 0$)	24	0.0013	0.0044
$J = 2$, indep constraints	4	0.0010	0.0018
$J = 2$, correlated constraints	4	0.0011	0.0019
$J = 2$, $\Sigma_{yg} \neq 0$	24	0.0007	0.0021
ICM-fitted posterior	200	0.0013	0.0039
Total	526	—	—

Table 3: Tail-probability verification at acquisition-relevant thresholds. $\Pr(\Delta > \varepsilon) = 1 - F_\Delta(\varepsilon)$ is the upper-tail probability the acquisition function consumes. For each case a single representative configuration is shown. Analytical values come from the analytical CDF; Monte Carlo values are linearly interpolated from the empirical CDF on the same δ-grid. Pointwise empirical-CDF standard error at $N = 100{,}000$ is ≈ 0.0016.

Case	ε	Analytical $\Pr(\Delta > \varepsilon)$	Monte Carlo	Abs. error
Unconstrained, indep	0	1.0000	1.0000	0.0000
	0.050	0.7110	0.7119	0.0009
	$q_{50} = 0.085$	0.5000	0.5012	0.0012
	$q_{90} = 0.143$	0.1000	0.1001	0.0001
	$q_{99} = 0.156$	0.0100	0.0099	0.0001
Unconstrained, corr	0	1.0000	1.0000	0.0000
	0.050	0.7520	0.7511	0.0009
	$q_{50} = 0.092$	0.5000	0.4989	0.0011
	$q_{90} = 0.145$	0.1000	0.0998	0.0002
	$q_{99} = 0.156$	0.0100	0.0099	0.0001
Constrained, indep	0	0.6441	0.6453	0.0012
	0.050	0.3998	0.4005	0.0007
	$q_{50} = 0.029$	0.5000	0.5009	0.0009
	$q_{90} = 0.158$	0.1000	0.1000	0.0000
Constrained, corr objs ($\Sigma_{yg} = 0$)	0	0.6192	0.6187	0.0005
	0.050	0.4087	0.4088	0.0001
	$q_{50} = 0.028$	0.5000	0.4998	0.0002
	$q_{90} = 0.203$	0.1000	0.1001	0.0001
Constrained, $\Sigma_{yg} \neq 0$, $J = 1$	0	0.3493	0.3501	0.0007
	0.050	0.2180	0.2188	0.0008
	$q_{50} = 0$	0.5000	0.5004	0.0004
	$q_{90} = 0.095$	0.1000	0.1008	0.0008
	$q_{99} = 0.326$	0.0100	0.0103	0.0003
$J = 2$ indep constraints	0	0.2675	0.2683	0.0008
	0.050	0.1623	0.1625	0.0002
	$q_{50} = 0$	0.5000	0.5011	0.0011
	$q_{90} = 0.080$	0.1000	0.0998	0.0002
	$q_{99} = 0.277$	0.0100	0.0100	0.0000
$J = 2$ correlated constraints	0	0.2119	0.2110	0.0010
	0.050	0.1286	0.1279	0.0007
	$q_{50} = 0$	0.5000	0.4994	0.0006
	$q_{90} = 0.067$	0.1000	0.0993	0.0007
	$q_{99} = 0.247$	0.0100	0.0096	0.0004

Continued on next page.

Continued from previous page.

Case	ε	Analytical $\Pr(\Delta > \varepsilon)$	Monte Carlo	Abs. error
$J = 2$, $\Sigma_{yg} \neq 0$	0	0.1484	0.1488	0.0004
	0.050	0.0939	0.0942	0.0003
	$q_{50} = 0$	0.5000	0.5002	0.0002
	$q_{90} = 0.044$	0.1000	0.1003	0.0003
	$q_{99} = 0.245$	0.0100	0.0105	0.0005
ICM-fitted posterior	0	0.0000	0.0000	0.0000
	0.050	0.0000	0.0000	0.0000
	$q_{50} = 0$	0.5000	0.4999	0.0001
	$q_{90} = 0$	0.1000	0.0998	0.0002
	$q_{99} = 0$	0.0100	0.0098	0.0002

Table 2 reports the grid-max CDF error per case. All 526 configurations pass the 0.01 threshold. The largest observed error is 0.0059, occurring in Case 2 (unconstrained correlated) on the sparse single-point Pareto front at $\sigma = 0.5$, $\rho = 0.9$. This exceeds the per-configuration DKW 95% band of $\sqrt{\log(2/0.05)/(2N)} \approx 0.0043$ but remains below the Bonferroni-adjusted family-wise band $\sqrt{\log(2K/0.05)/(2N)} \approx 0.0071$ for $K = 526$ configurations. Across the full grid, 6 configurations exceed the per-configuration band and no configuration exceeds the family-wise band. Because the DKW band is a conservative upper bound rather than a calibrated null, these exceedance counts are diagnostic only: the observed deviations are compatible with empirical-CDF sampling variability and remain far below the predefined 0.01 acceptance threshold. Per-case mean grid-max errors sit in the 7×10^{-4} to 2×10^{-3} range (Table 2, "Mean grid-max err" column), consistent with the $\mathcal{O}(1/\sqrt{N}) \approx 10^{-3}$ empirical-CDF sampling scale.

In addition to the per-CDF verification, Table 3 reports analytical and Monte Carlo *tail* probabilities $\Pr(\Delta > \varepsilon)$ at acquisition-relevant thresholds: $\varepsilon = 0$ (the threshold at which the acquisition value reduces to the constrained PoHVI of Wang et al.), $\varepsilon = 0.05$ (a typical mid-range value of the BO loop's ε_t schedule), and the 50th, 90th and 99th percentiles of the analytical Δ distribution. This is the quantity the ε-PoHVI/ε-CPoHVI acquisition function actually consumes. Across every case the analytical and Monte Carlo tail probabilities agree to within 0.0012; the worst disagreement is on the simplest constrained case (Case 3) at $\varepsilon = 0$ rather than on the analytically newest cases. That this is the right qualitative behavior: cases with stronger feasibility-induced concentration of mass (the $\Sigma_{yg} \neq 0$ and $J = 2$ blocks) push the bulk of the distribution away from $\varepsilon = 0$ and so leave less empirical-CDF mass near the threshold to fluctuate, whereas the simpler cases place more mass in that neighbourhood and therefore have more room for raw sampling variability. Tail-probability errors are uniformly smaller than the CDF grid-max errors of Table 2, as expected, since the empirical-CDF deviation upper-bounds the pointwise tail-probability error, so any downstream BO sensitivity to misranking by the acquisition function is bounded above by the per-case grid-max error rather than by the (typically larger) raw $|F_{\mathrm{MC}} - F|$ pointwise gap.

Figure 1 shows the unconstrained verification at three representative correlation levels ($\rho \in \{-0.9, 0, +0.9\}$) on a fixed `moderate`/`near_front`/$\sigma = 0.2$ configuration. The top row overlays the analytical and Monte Carlo CDFs (the curves visually overlap to within line thickness); the bottom row plots the residual $\widehat{F}_N - F$ with the per-configuration DKW 95% band shaded. The residual stays inside the band on every panel. Figure 2 plots the full mixed CDF $F_{\Delta_C}(\delta)$ for the three

feasibility regimes `easy_feasible`, `borderline`, and `tight`; the visible vertical jump at $\delta = 0$ is the infeasible atom of size $1 - \mathrm{PoF}_{\mathrm{joint}}$. Figure 3 reports Case 8 as a per-problem boxplot of the grid-max error plus a worst-case residual panel for each problem (with the DKW band overlay). Figure 4 aggregates the unconstrained Cases 1 and 2 by averaging the grid-max error over (μ, PF) configurations within each (ρ, σ) cell. Figure 5 shows the violin plot of the grid-max error across all 526 configurations grouped by case.

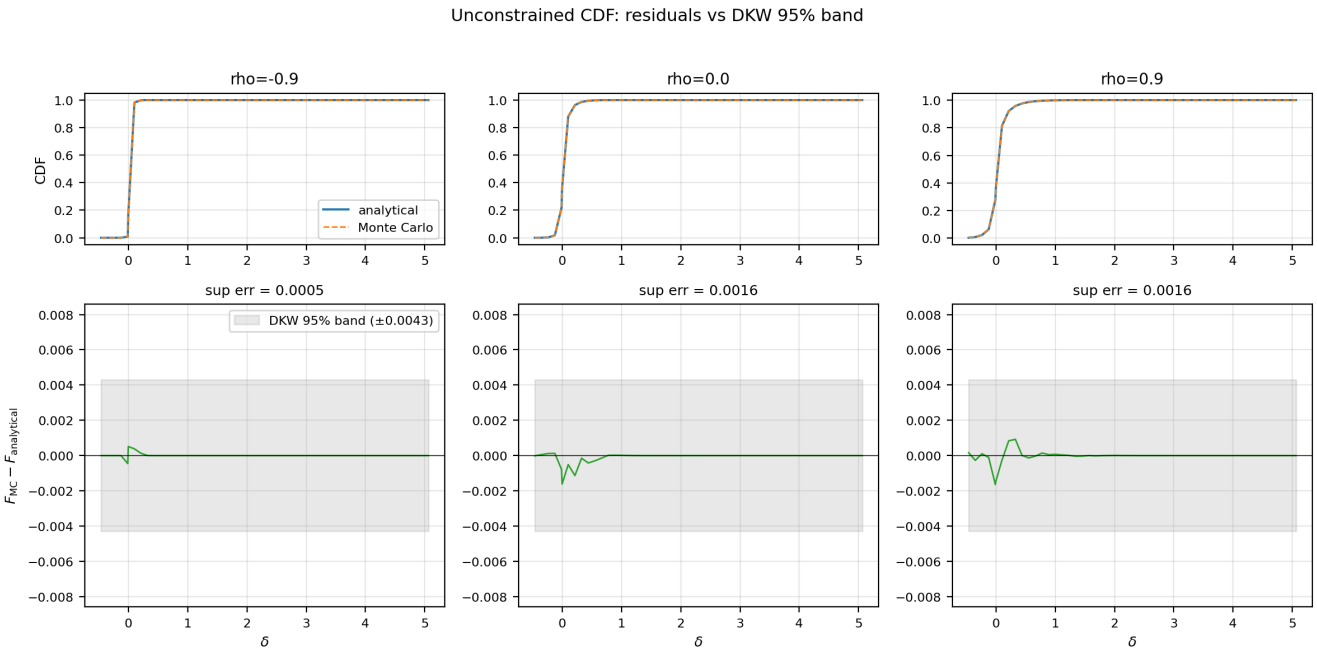

Figure 1: Unconstrained verification. *Top:* analytical (solid) and Monte Carlo (dashed) CDFs for the same $(\mathrm{PF}, \mu, \sigma) = (\texttt{moderate}, \texttt{near\_front}, 0.2)$ configuration at three correlation levels. *Bottom:* residual $\widehat{F}_N - F$ with the per-configuration DKW 95% band shaded (± 0.0043 at $N = 10^5$). The residual is inside the band on every panel.

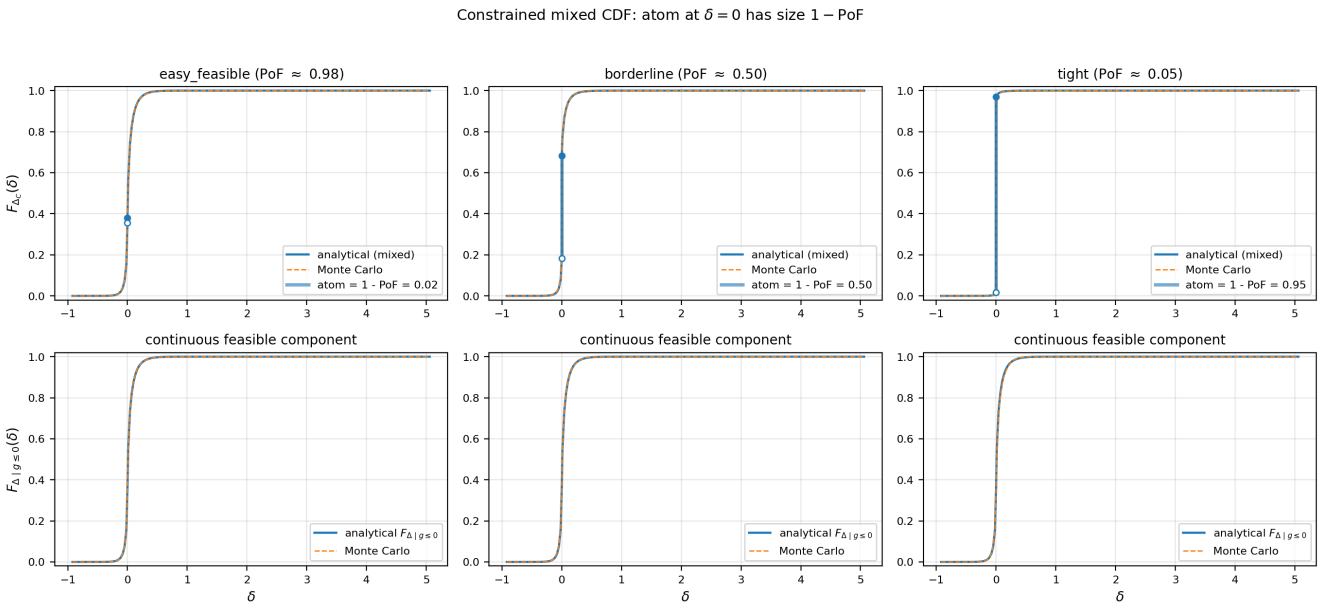

Figure 2: Constrained mixed CDF for the three feasibility regimes (Cases 3 and 4). *Top row:* full mixed CDF $F_{\Delta_C}(\delta)$. The vertical jump at $\delta = 0$ is the infeasible point mass of size $1 - \mathrm{PoF}_{\mathrm{joint}}$ (filled vs. open marker indicates the inclusive end of the jump). *Bottom row:* continuous feasible component $F_{\Delta|g\leq 0}(\delta)$, which under independence reduces to the marginal $F_{\mathrm{HVI}}(\delta)$. Plotting the feasible-conditional component separately keeps it visible in the tight-feasibility regime, where the atom otherwise dominates. Analytical (solid) and Monte Carlo (dashed) overlap to within line thickness in each panel.

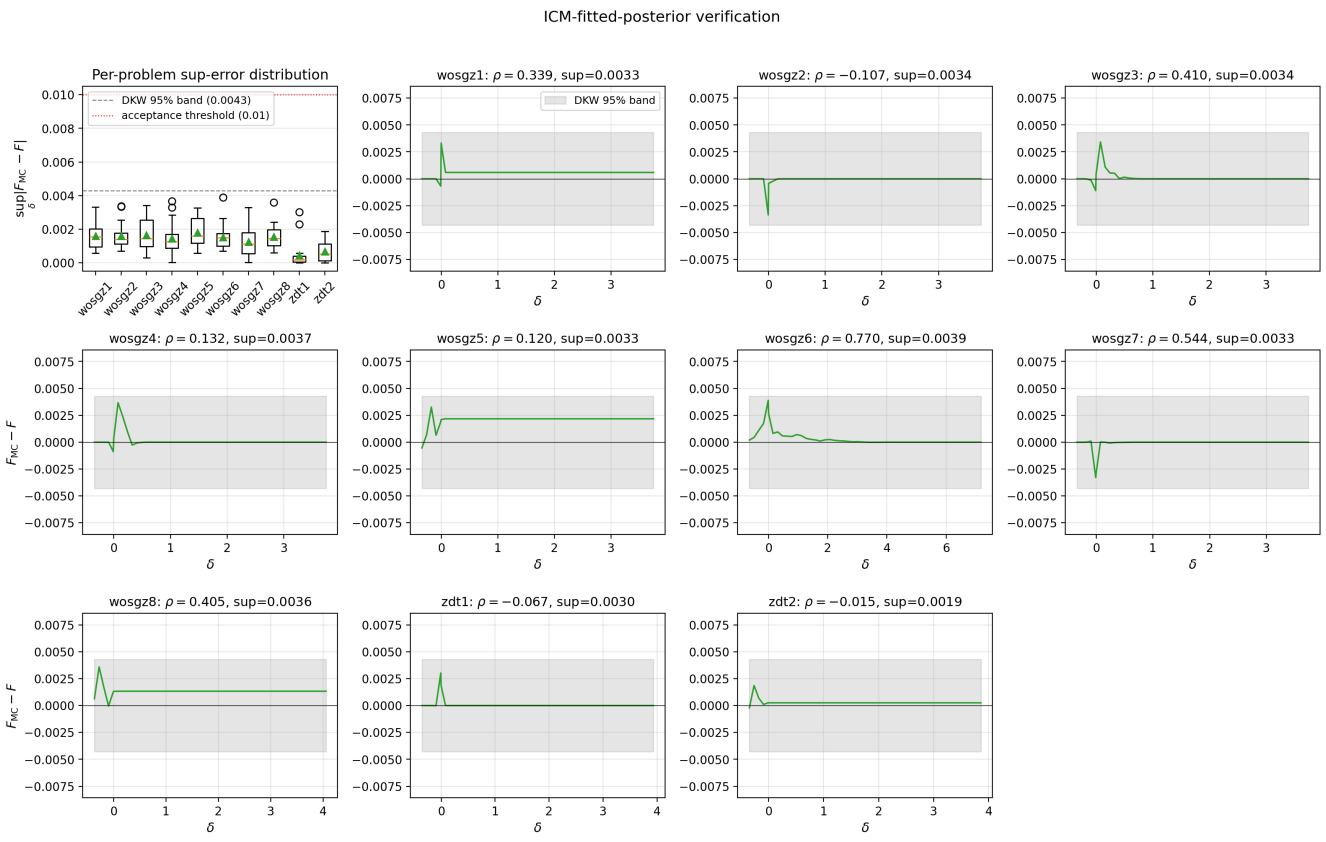

Figure 3: ICM-fitted-posterior verification (Case 8). *Top-left:* per-problem boxplot of the grid-max CDF error across the 20 candidate-point posteriors per problem (ZDT1, ZDT2, WOSGZ1 to WOSGZ8). *Remaining cells:* residual $\widehat{F}_N - F$ for the worst-case configuration in each problem, with the DKW 95% band overlaid. Every configuration passes the 0.01 acceptance threshold. Each panel title gives that panel's ρ and grid-max error.

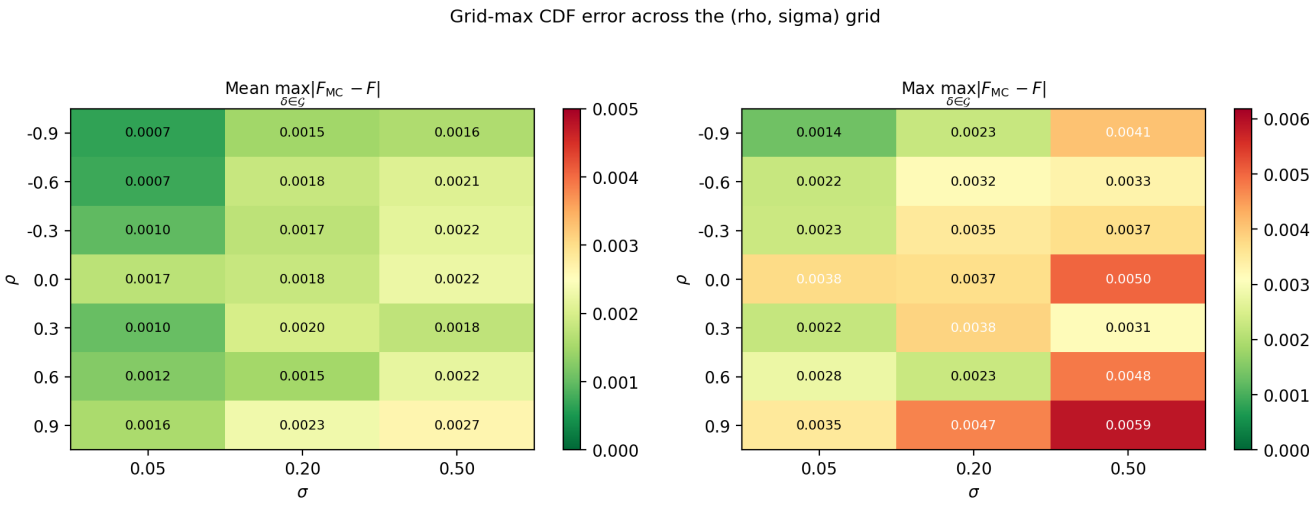

Figure 4: Grid-max CDF error for Cases 1 and 2 across the (ρ, σ) parameter grid. *Left:* mean over the 12 (μ, PF) configurations within each cell. *Right:* maximum over the same configurations. Both panels auto-scale their colorbar to the observed range; errors remain at the Monte Carlo noise scale across the tested grid. Pairing mean and max in the same figure shows that no (ρ, σ) cell hides a localized failure under a benign mean.

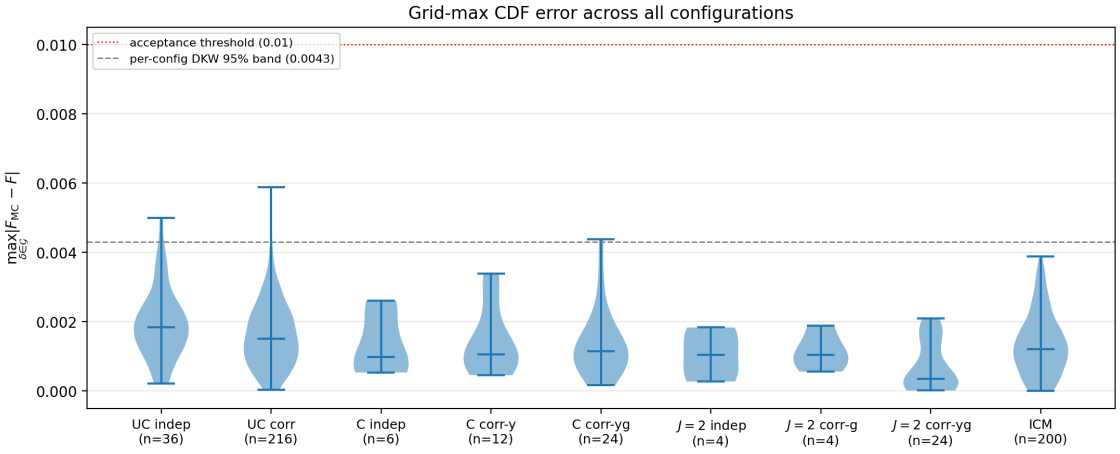

Figure 5: Distribution of the grid-max CDF error across the full 526-configuration verification grid, grouped by case (UC = unconstrained, C = constrained, $J = 2$ = two-constraint variants, ICM = realistic-posterior). Each violin shows the per-configuration spread, not just the mean. All configurations lie far below the 0.01 acceptance threshold (red dotted); most lie within the per-configuration DKW 95% band (gray dashed) of 0.0043; six configurations sit above this conservative band, and none exceeds the Bonferroni-adjusted family-wise band of 0.0071 at $K = 526$.

Deterministic sanity checks. Alongside the Monte Carlo comparison, the verification harness checks identities that the analytical implementation must satisfy regardless of sampling. Table 4 reports the worst deviations across the full grid; the first five rows are deterministic identities that

should hold up to quadrature and floating-point tolerance, and the last is statistical (carrying Monte Carlo noise rather than numerical tolerance). The deterministic rows confirm that the cell-partition probabilities sum to one, that the analytical CDFs are monotone non-decreasing, and that the correlated and constrained-correlated code paths reduce to their simpler counterparts when the relevant correlation blocks are set to zero. The two-order-of-magnitude gap between the $\rho = 0$ reduction (1.4×10^{-5}) and the two $\Sigma_{yg} = 0$ reductions ($\sim!10^{-7}$) is itself informative: the $\rho = 0$ check compares the *correlated* code path (Drezner-Wesolowsky bivariate-CDF quadrature feeding the within-cell density) against the *independent* code path (analytical $\Phi(a)\Phi(b)$ factorization), so the residual reflects accumulated quadrature and floating-point error from two structurally distinct implementations evaluating the same mathematical object. The $\Sigma_{yg} = 0$ reductions stay within the constrained-correlated code path with one covariance block zeroed, so the residual is governed only by the conditioning arithmetic and is correspondingly smaller.

Table 4: Deterministic and Monte Carlo sanity checks across the full verification grid. "Deterministic" checks should hold up to quadrature and floating-point tolerance; "MC noise" checks are bounded by the empirical-CDF standard error at $N = 10^5$.

Check	Worst deviation	Type
Cell probabilities sum to one	1.6×10^{-8}	deterministic
Analytical CDF monotonicity violation[†]	1.4×10^{-7}	deterministic
$\rho = 0$ correlated reduces to independent	1.4×10^{-5}	deterministic
$\Sigma_{yg} = 0$ corr-yg reduces to Case 3 / Case 4	6.6×10^{-7}	deterministic (Case 5 reduction)
$\Sigma_{yg} = 0$ in $J = 2$ corr-yg reduces to Case 7	1.7×10^{-7}	deterministic (Case 9 reduction)
Empirical infeasible mass $\widehat{P}(g > 0)$ vs. $1 - \mathrm{PoF}_{\mathrm{joint}}$	8×10^{-4}	MC noise

[†] The analytical CDF is mathematically monotone non-decreasing; the reported deviation is accumulated quadrature and floating-point error in the cell-density integration, not a mathematical violation.

6.3.3 Analysis

First, the unconstrained correlated implementation shows no systematic deterioration over the tested correlation range. The single largest synthetic-grid error occurs at $\rho = 0.9$ on the sparse single-point Pareto front at $\sigma = 0.5$, but the paired mean/max heatmap in Figure 4 does not show monotone growth in $|\rho|$; errors remain within the empirical-CDF sampling scale across $\rho \in [-0.9, +0.9]$. The corner where the worst case sits combines the stress factors that the cell-density quadrature is most sensitive to: a single-point Pareto front maximizes the number of cells whose mass depends on the bivariate-normal rectangle probability rather than its univariate factorisation, a wide marginal posterior spreads mass across many of those cells, and $|\rho| = 0.9$ pushes the within-cell density toward the singular axis where the Laurent coefficients of Proposition 2.33 are largest. Even there the deviation remains below the family-wise Bonferroni band.

Second, the constrained cases sit at the same error scale as the unconstrained ones. The single-constraint $\Sigma_{yg} \neq 0$ path (Case 5) reaches a maximum grid-max error of 0.0044; the $J = 2$ joint-feasibility orthant path (Case 7) reaches 0.0019; and the full $J = 2$, $\Sigma_{yg} \neq 0$, $\Sigma_{gg} \neq 0$ multivariate-conditioning path (Case 9) reaches 0.0021. These are the cases that exercise the analytically newest pieces of the constrained derivation; their staying at the same noise scale as the simpler product-weighted cases is evidence that the outer-integral, orthant-probability, and

tensor-product Gauss-Legendre quadrature implementations agree with the Monte Carlo reference on the tested configurations.

Third, the fitted-posterior block verifies the CDF implementation on the posterior parameter regimes the selected ICM surrogate actually produces. Across all 200 ICM-fitted posteriors on ZDT1, ZDT2 and WOSGZ1-WOSGZ8 the worst grid-max error is 0.0039, occurring on a WOSGZ6 candidate point at $\rho \approx 0.77$. The per-problem boxplot in Figure 3 shows uniform error scales across all ten problem blocks, including the WOSGZ4-WOSGZ8 problems on which Chapter 8 reports the strongest empirical wins. At $N = 10^5$ samples the verification is bounded by Monte Carlo sampling noise: any residual analytical error sits below the resolution at which the empirical CDF can detect it, so tightening the verification would require an order-of-magnitude larger Monte Carlo budget rather than any change to the analytical machinery.

Coverage summary. The grid exercises every distributional code path used by the thesis: unconstrained independent and correlated objectives (Cases 1 and 2, with $\rho \in -0.9, -0.6, -0.3, 0, +0.3, +0.6, +0.9$); constrained HVI with $\Sigma_{yg} = 0$ for $J = 1$ (Cases 3 and 4) and $J = 2$ (Cases 6 and 7); constrained HVI with $\Sigma_{yg} \neq 0$ for $J = 1$ across feasibility regimes from PoF ≈ 0.05 to 0.98 (Case 5); the joint-feasibility orthant probability with both independent and correlated constraints (Cases 6 and 7); and the full $J = 2$, $\Sigma_{yg} \neq 0$, $\Sigma_{gg} \neq 0$ multivariate-conditioning path (Case 9). The deterministic reduction checks of Table 4 confirm that setting the relevant correlation blocks to zero recovers the corresponding simpler implementation to within quadrature tolerance.

Constrained acquisition: scope for the BO benchmark. The ε-CPoHVI distribution (Chapter 4) is verified above against Monte Carlo, but is not benchmarked in a full BO loop in this thesis. The natural future-work targets are the standard constrained multi-objective acquisition baselines, the constrained expected hypervolume improvement of Feliot et al. [13], the noisy parallel variants NEHVI-C / qNEHVI-C [8], and the entropy-based MESMOC [4] and PF^{2}ES [24], benchmarked on the established constrained MOO test suites C1/C2/C3-DTLZ and the MW1-MW14 suite of Ma and Wang. A fair comparison across these methods requires a separate experimental study, because the baselines differ in surrogate assumptions, feasibility models, noise handling, and batch support. This thesis therefore stops at the analytical and numerical foundation of ε-CPoHVI and leaves the full constrained BO comparison to that follow-up study.

7 Surrogate Model Selection

The correlated ε-PoHVI acquisition function consumes the full bivariate predictive distribution $\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x}))$ of an underlying surrogate model. Independent Gaussian processes, by construction, produce $\rho(\boldsymbol{x}) \equiv 0$ and therefore cannot supply the cross-objective signal the correlated acquisition is built to exploit. A multi-output Gaussian process (MOGP) is required, and the choice between competing MOGP pipelines is an empirical question. This chapter benchmarks five surrogate pipelines, one independent-GP baseline plus four MOGPs, on the WOSGZ family and selects the one used in the BO experiment of Section 8.

7.1 Setup

Four MOGP implementations are benchmarked against the independent-GP baseline on **WOSGZ1, WOSGZ2, and WOSGZ3** at $d = 30$. Benchmarking across three problems rather than one addresses the sample-of-one-at-the-problem-level concern: a selection decision based on a single problem may fail to generalize. WOSGZ problems also avoid ZDT1's artifact that $f_1 = x_1$ is trivially predictable; on WOSGZ both objectives carry non-trivial modeling burden via the 15-variable distance terms, so the RMSE ratio on both objectives is meaningful.

Training set sizes are $n \in \{60, 120, 230\}$ matching the BO-loop regime (initial sample, one doubling, and the terminal training size after 170 acquisition iterations) with a held-out test set of $N_{\text{test}} = 200$ points per fit. All training and test inputs are drawn by Latin Hypercube sampling over the problem's box and evaluated with the thesis's own WOSGZ implementation. Each of the five model configurations is fitted on the same training set per seed, and predictions are scored against the same test set, so every baseline-vs-MOGP and MOGP-vs-MOGP comparison is paired by seed. Ten seeds are used. The total budget is 5 models $\times$ 3 problems $\times$ 3 n-values $\times$ 10 seeds = 450 fits.

7.1.1 Candidate surrogate pipelines

The five candidates differ in more than just the multi-output GP architecture: they differ in the base kernel (Matérn 3/2 for ICM and LMC, BoTorch's default RBF for the unmodified MTGP path, spectral mixture for SM-LMC), in the optimizer (Adam for the GPyTorch and MOGPTK MOGPs, L-BFGS-B for the independent baseline and BoTorch MTGP), in the restart budget, and in the underlying library. The benchmark therefore selects a surrogate pipeline for use with the correlated acquisition.

All MOGP candidates produce a full 2×2 posterior covariance per test point, from which $\rho(\boldsymbol{x}) = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$ is extracted for use by the correlated acquisition function. The baseline only produces a diagonal covariance.

Independent GP (baseline). One `sklearn GaussianProcessRegressor` per objective with kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \theta_0 \cdot k_{\text{Mat}}(\boldsymbol{x}, \boldsymbol{x}'; \nu, \boldsymbol{\ell}), \tag{143}$$

where $\theta_0 \in [10^{-5}, 10^3]$ is the signal variance, $\nu = \frac{3}{2}$ (Matérn 3/2), and $\boldsymbol{\ell} \in [10^{-10}, 10^5]^d$ is the anisotropic length-scale vector. Hyperparameters are optimized by maximizing the log-marginal likelihood via L-BFGS-B with $5d = 150$ random restarts. Training targets are standardized and a nugget $\alpha = 10^{-5}$ is added for numerical stability.

The choice of $\nu = \frac{3}{2}$ rather than the $\nu = \frac{5}{2}$ reported in the paper text of Wang et al. [27] is deliberate: the released code's `args.yml` archive sets `nu: 3`, which the codebase maps to $\nu = 0.5 \times 3 = 1.5$, and the Wang et al. [27] baseline runs that this thesis compares against were produced with the same setting.

GPyTorch ICM, rank 1. Intrinsic Coregionalization Model [2, 6], fitted with GPyTorch's `MultitaskKernel`. Matérn 3/2 base kernel with ARD length scales $\boldsymbol{\ell} \in \mathbb{R}^{30}$, a scale kernel, and a task coregionalization matrix $B = WW^\top + \mathrm{diag}(\boldsymbol{v})$ with $W \in \mathbb{R}^{2\times 1}$ and $\boldsymbol{v} \in \mathbb{R}^2$. Hyperparameters are optimized by Adam (learning rate 0.1, 1000 iterations) with 3 random restarts; the noise variance is lower-bounded at 10^{-4} for numerical stability. The model exposes $R_{12} = B_{12}/\sqrt{B_{11}B_{22}}$ as a scalar task-correlation estimate.

Rank 1 is sufficient for two-task task-covariance expressiveness. Any 2×2 symmetric PSD matrix B satisfies $B_{12}^2 \leq B_{11}B_{22}$, which is exactly the condition under which there exist $w \in \mathbb{R}^2$ and $\boldsymbol{v} \geq \mathbf{0}$ such that $B = ww^\top + \mathrm{diag}(\boldsymbol{v})$. Rank 1 + diagonal therefore spans the full 2-task PSD cone, and rank 2 would add two more parameters without adding expressive power. Rank 1 is also better regularized at the small training-set sizes used in the BO loop.

GPyTorch LMC, rank 2. Linear Model of Coregionalization [2] with two latent GPs, each with its own Matérn 3/2 base kernel (ARD) and scale kernel. The task assignments are learned per latent GP, so the model is strictly more flexible than rank-1 ICM at the cost of $O(2 \cdot d)$ extra length-scale parameters. Same optimizer as ICM (Adam, 2000 iterations, 3 restarts). An effective R_{12} is computed from the learned latent-task-assignment matrices and reported.

BoTorch MultiTaskGP. BoTorch's `MultiTaskGP` [3], an ICM model with an RBF base kernel (the BoTorch default) and task rank 2. Fitted with `fit_gpytorch_mll` (L-BFGS-B, up to 2 retries, 3 restarts). Because BoTorch's default `PositiveIndexKernel` constrains $B_{mm'} \geq 0$ and WOSGZ can induce negative learned task correlations, the task kernel is replaced with GPyTorch's `IndexKernel`, preserving the `active_dims` argument of the original kernel.

MOGPTK SM-LMC. A spectral-mixture LMC [9] with $Q = 1$ spectral component and latent-rank 1, fitted for 2000 iterations with 3 restarts. The SM kernel is intended for data with periodic or non-stationary structure, which is a poor match for the smooth bias-and-distance form of WOSGZ. SM-LMC is included as a prior-mismatched control: this $Q = 1$ SM-LMC configuration is expected to be empirically mismatched to WOSGZ under the present protocol, and is benchmarked to confirm the expectation under the same protocol as the viable candidates.

Iteration and restart budgets. Per-model Adam iteration counts (ICM rank 1: 1000; LMC rank 2 and SM-LMC: 2000) and restart counts (3 random restarts per MOGP, $5d = 150$ L-BFGS-B restarts for the independent baseline) follow from two pilot studies reported in Appendix A. The asymmetric restart counts (150 for the baseline, 3 for every MOGP) are deliberately tilted in favor of the baseline, so any MOGP win under this configuration is a conservative result; a sensitivity check at ICM rank 1 confirms that the 3-restart budget is not a limiting factor for the selected architecture.

7.1.2 Metrics

Six metrics are reported per (model, problem, n, seed) and aggregated over seeds:

- **Marginal NLPD.** Per-objective negative log predictive density, averaged over the 200 test points, with the posterior marginal standard deviation $\sigma_k(\boldsymbol{x}^*) = \sqrt{\Sigma_{kk}(\boldsymbol{x}^*)}$ used as the Gaussian width. This is the standard one-dimensional NLPD; it measures the combined quality of the predictive mean $\mu_k(\boldsymbol{x}^*)$ and marginal uncertainty $\sigma_k(\boldsymbol{x}^*)$ but does *not* use Σ_{12}. See "joint vs. marginal NLPD" below for how this metric relates to the acquisition's full-$\boldsymbol{\Sigma}$ input.

- **coverage$_{95}$.** Fraction of test points whose true $\boldsymbol{y}$ lies inside the 95% credible ellipse implied by the predicted $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Computed from the Mahalanobis distance $d^2 = \boldsymbol{r}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{r}$ (with $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{\mu}$) and the $\chi^2(2)$ 95th-percentile threshold 5.991. Unlike marginal NLPD, this metric uses the full 2×2 posterior covariance through $\boldsymbol{\Sigma}^{-1}$, so it directly probes the joint calibration the correlated acquisition relies on. Target value: 0.95; under- or over-coverage are both calibration failures.

- **RMSE ratio** (f_1, f_2)**.** Prediction RMSE divided by the independent-GP baseline's RMSE per objective. A ratio < 1 means the MOGP predicts that objective more accurately than the baseline. Used as a check on mean-prediction accuracy independent of uncertainty calibration.

- $|\bar{\rho}|$ **(mean absolute posterior correlation).** Mean of $|\rho(\boldsymbol{x}^*)|$ over the 200 test points, where $\rho(\boldsymbol{x}^*) = \Sigma_{12}(\boldsymbol{x}^*)/\sqrt{\Sigma_{11}\Sigma_{22}}$ is extracted from the posterior covariance. This is the per-point cross-objective signal that the correlated acquisition exploits; $|\bar{\rho}| = 0$ by construction for the baseline.

- R_{12} **(learned task correlation).** Scalar $B_{12}/\sqrt{B_{11}B_{22}}$, global across the input space, for models that expose it.

- **MACE-ρ (diagnostic).** Mean absolute calibration error of the predicted cross-covariance. Test points are binned into five quintiles by predicted $\rho(\boldsymbol{x}^*)$; within each quintile the empirical Pearson correlation of the marginal-standardized residuals $(y_k - \mu_k)/\sqrt{\Sigma_{kk}}$ across objectives $k = 1, 2$ is computed, and the absolute discrepancy between per-quintile mean predicted ρ and per-quintile empirical correlation is averaged, weighted by bin size. This is reported as a diagnostic, not as a primary selection criterion: with ~ 40 test points per quintile the per-bin empirical Pearson correlations carry standard error $\approx 1/\sqrt{40} \approx 0.16$, so MACE-$\rho$ values below that scale are within the resolution of the diagnostic. For the baseline (diagonal $\boldsymbol{\Sigma}$, all predicted $\rho = 0$) the quintile machinery degenerates; $|\rho_{\text{emp}}|$ over all 200 test points is reported directly, framed as "cross-objective structure the baseline leaves on the table". This baseline value is a different quantity from the MOGP MACE-ρ and the two are not directly comparable; the relevant cross-correlation comparison is made through coverage$_{95}$, which uses both implementations on a common scale.

Joint vs. marginal NLPD: what the co-primary pairing actually measures. The correlated acquisition consumes the full bivariate predictive $\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x}))$ at each candidate, so the

formally aligned scalar metric is the *joint* NLPD,

$$\mathrm{NLPD}_{\mathrm{joint}}(\boldsymbol{x}^*) = -\log \phi_2\big(\boldsymbol{y};\, \boldsymbol{\mu}(\boldsymbol{x}^*),\, \boldsymbol{\Sigma}(\boldsymbol{x}^*)\big) = \log(2\pi) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}| + \tfrac{1}{2}\boldsymbol{r}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{r},$$

which uses Σ_{12} explicitly. The marginal-NLPD metric reported here, $\mathrm{NLPD}_{\mathrm{marg}} := \mathrm{NLPD}_{f_1} + \mathrm{NLPD}_{f_2}$, decomposes the joint as

$$\mathrm{NLPD}_{\mathrm{joint}} = \mathrm{NLPD}_{\mathrm{marg}} + \underbrace{\tfrac{1}{2}\log(1-\rho^2)}_{\text{volume term}} + \underbrace{\frac{\rho^2(z_1^2+z_2^2)-2\rho z_1 z_2}{2(1-\rho^2)}}_{\text{coupling term}},$$

where $z_k = (y_k - \mu_k)/\sigma_k$ and $\rho = \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$. The volume term is non-positive: predicting any nonzero ρ shrinks the predictive volume relative to the independent-marginal product, so the log-determinant contribution to NLPD is reduced. The coupling term is bounded in magnitude by $|\rho|/(1-\rho^2)$ times a quantity of $\mathcal{O}(z_k^2)$, and is small whenever $|\rho|$ is small. For the baseline ($\rho \equiv 0$) both terms are zero, so $\mathrm{NLPD}_{\mathrm{marg}} = \mathrm{NLPD}_{\mathrm{joint}}$ exactly. For an MOGP with predicted $|\bar{\rho}|$ in the 0.06-0.27 range observed in this benchmark (Table 12), the joint-marginal gap is small relative to the marginal-NLPD differences between models, so ranking on marginal NLPD aligns with ranking on joint NLPD up to a correction term whose sign and magnitude is bounded by the predicted $|\bar{\rho}|$. The joint-calibration component the marginal NLPD misses, whether $\rho(\boldsymbol{x}^*)$ is the right shape, not just $|\rho(\boldsymbol{x}^*)|$ being nonzero, is captured separately by $\mathrm{coverage}_{95}$, which uses $\boldsymbol{\Sigma}^{-1}$ directly.

Marginal NLPD and $\mathrm{coverage}_{95}$ are therefore treated as co-primary metrics: a model that wins on both wins on a quantity at least as strict as joint NLPD up to a small correction, while a model that wins on only one is flagged for further scrutiny.

Selection metric. Selection is grounded in the co-primary pair (marginal NLPD, $\mathrm{coverage}_{95}$), with $|\bar{\rho}|$ and MACE-ρ reported as diagnostics and RMSE as an independent cross-check on mean-prediction accuracy. This pairing addresses the gap between "what the acquisition consumes" (the full bivariate density) and "what a tractable per-objective scalar metric measures" (marginal calibration only).

7.1.3 Statistical methodology

Descriptive means over ten seeds do not quantify whether the apparent ranking is statistically meaningful. This experiment complements the descriptive numbers with paired non-parametric tests, multiple-comparison correction, and non-parametric effect-size measures.

Wilcoxon signed-rank. For baseline-vs-MOGP comparisons on RMSE and NLPD the test is $H_0 : \mathrm{metric}_{\mathrm{MOGP}} \geq \mathrm{metric}_{\mathrm{baseline}}$ (one-sided, paired by seed, rejecting towards "MOGP is better"), using the Wilcoxon signed-rank test [29]. Each seed produces the same LHS training and test set for every model, so the paired design is exact. The non-parametric form avoids an unwarranted normality assumption on the per-seed differences given the sample size of ten. Pairwise MOGP-vs-MOGP comparisons use the two-sided variant.

Holm-Bonferroni correction. Nine families of hypotheses are pre-declared, each corrected separately by the Holm-Bonferroni step-down procedure [18]:

- **Baseline-vs-MOGP, pooled across problems** (RMSE): 4 MOGPs $\times$ 2 objectives $\times$ 3 n-values = 24 tests.
- **Baseline-vs-MOGP, pooled across problems** (NLPD): 24 tests on the marginal-NLPD metric, corrected separately.
- **Per-problem breakdown**: 3 problems $\times$ 4 MOGPs = 12 tests, computed separately for each of RMSE-f_2, RMSE-f_1, NLPD-f_2, NLPD-f_1 (four 12-test families, 48 tests in total).
- **Pairwise MOGP-vs-MOGP at** $n = 230$, RMSE: $\binom{4}{2} \times 2 = 12$ two-sided tests.
- **Pairwise viable-MOGP-vs-MOGP at** $n = 230$ on the co-primary metrics (Table 11): 3 viable-MOGP pairs corrected within each of marginal NLPD and coverage$_{95}$, giving two further families of 3 tests each.

The total is $24 + 24 + 48 + 12 + 3 + 3 = 114$ tests across the nine families. Like Bonferroni, Holm controls the family-wise error rate under arbitrary dependence, but it is uniformly less conservative than single-step Bonferroni; this combination of Wilcoxon + Holm is a standard recipe for comparing learning algorithms across shared training sets [10]. The families are corrected separately rather than as one combined family because they address distinct questions, and combining them would impose a correction conservative enough to collapse the small-sample power.

A caveat on within-family pooling. Several of the per-cell tests below pool across the three n-values or the three problems (e.g. a per-problem NLPD cell pools 10 seeds $\times$ 3 n-values = 30 paired differences). The three n-values within a (problem, seed) triplet share the same Latin Hypercube design template, so the pooled differences are not 30 independent observations; the effective sample size is between 10 and 30. The Wilcoxon p-values reported below should therefore be read as descriptive inferential summaries rather than fully calibrated significance tests; if the pooled differences are positively correlated within seed/problem blocks (as is plausible here), the nominal p-values may be optimistic, and they are read in conjunction with effect-size indicators (median difference, win rate) rather than as stand-alone hypothesis tests.

Win rate. As a non-parametric effect-size complement, the per-seed win count is reported for each test. When the table pools over both problems and n-values within a (model, n or problem)-cell, the denominator is 10 seeds $\times$ 3 pooled levels = 30. When the table pools only over problems within an n-cell the denominator is $10 \times 3 = 30$ as well. When the table pools only over n-values within a (problem, model)-cell the denominator is $10 \times 3 = 30$. The denominator is annotated in each table caption. A low Holm-corrected p-value with a high win rate is the strongest evidence pattern.

7.2 Results

Results are presented in the order that mirrors the selection-metric argument of Section 7.1.2: NLPD first, RMSE second, calibration third, pairwise MOGP comparisons, then a descriptive reference table. Significance markers throughout: * $p_{\text{Holm}} < 0.05$, ** $p_{\text{Holm}} < 0.01$, *** $p_{\text{Holm}} < 0.001$.

7.2.1 Marginal NLPD

Tables 5 and 6 report per-problem one-sided Wilcoxon tests on NLPD, pooling n-values by seed (so each (problem, model) cell pools 30 paired differences) with Holm-Bonferroni correction across the 12 tests within each table. Median differences are in NLPD units; negative values favor the MOGP.

Table 5: Per-problem Wilcoxon signed-rank on NLPD-f_2 (one-sided; H_0: MOGP NLPD $\geq$ baseline NLPD). Holm correction across 3 problems $\times$ 4 MOGPs = 12 tests. Win count out of 10 seeds $\times$3 n-values = 30 matchups.

Problem	Model	p_{raw}	p_{Holm}	Win rate	Med. diff.
WOSGZ1	GPyTorch ICM (R=1)	<0.001	<0.001***	27/30	−0.528
WOSGZ1	GPyTorch LMC (R=2)	<0.001	<0.001***	26/30	−0.257
WOSGZ1	BoTorch MultiTaskGP	<0.001	<0.001***	24/30	−0.316
WOSGZ1	MOGPTK SM-LMC	1.000	1.000	5/30	+2.382
WOSGZ2	GPyTorch ICM (R=1)	<0.001	<0.001***	27/30	−0.563
WOSGZ2	GPyTorch LMC (R=2)	0.007	0.041*	19/30	−0.230
WOSGZ2	BoTorch MultiTaskGP	0.002	0.011*	21/30	−0.382
WOSGZ2	MOGPTK SM-LMC	1.000	1.000	7/30	+1.477
WOSGZ3	GPyTorch ICM (R=1)	<0.001	<0.001***	27/30	−0.248
WOSGZ3	GPyTorch LMC (R=2)	0.013	0.066	23/30	−0.106
WOSGZ3	BoTorch MultiTaskGP	0.013	0.066	21/30	−0.130
WOSGZ3	MOGPTK SM-LMC	1.000	1.000	4/30	+1.872

Table 6: Per-problem Wilcoxon signed-rank on NLPD-f_1; same structure as Table 5.

Problem	Model	p_{raw}	p_{Holm}	Win rate	Med. diff.
WOSGZ1	GPyTorch ICM (R=1)	<0.001	<0.001***	27/30	−0.450
WOSGZ1	GPyTorch LMC (R=2)	0.001	0.005**	24/30	−0.165
WOSGZ1	BoTorch MultiTaskGP	0.004	0.014*	19/30	−0.226
WOSGZ1	MOGPTK SM-LMC	1.000	1.000	6/30	+1.487
WOSGZ2	GPyTorch ICM (R=1)	<0.001	<0.001***	29/30	−0.611
WOSGZ2	GPyTorch LMC (R=2)	<0.001	<0.001***	25/30	−0.259
WOSGZ2	BoTorch MultiTaskGP	<0.001	<0.001***	23/30	−0.310
WOSGZ2	MOGPTK SM-LMC	1.000	1.000	8/30	+1.338
WOSGZ3	GPyTorch ICM (R=1)	<0.001	<0.001***	27/30	−0.384
WOSGZ3	GPyTorch LMC (R=2)	<0.001	<0.001***	25/30	−0.149
WOSGZ3	BoTorch MultiTaskGP	0.002	0.010*	22/30	−0.255
WOSGZ3	MOGPTK SM-LMC	1.000	1.000	6/30	+1.703

ICM rank 1 reaches $p_{\text{Holm}} < 0.001$ on all six (problem, objective) cells, with median NLPD reductions of -0.25 to -0.61 and win rates of 27-29 out of 30, after pooling across the three training sizes. This says ICM rank 1 outperforms the baseline on each (problem, objective) cell when the three n-values are treated together; it does not by itself certify significance at every individual n

within each cell, since the per-n samples are not independent (see the within-family pooling caveat in §7.1.3). The descriptive-table values in Table 12 confirm the per-n direction: ICM's mean NLPD beats the baseline at every (problem, n) cell (positive median differences in Tables 5-6 correspond to MOGP $\geq$ baseline NLPD, which the MOGP wants small). LMC and BoTorch MTGP reach Holm-significance on most cells but with larger p-values and smaller median effects. SM-LMC has larger NLPD than the baseline on every cell.

7.2.2 RMSE pooled across problems

Table 7 reports baseline-vs-MOGP Wilcoxon tests on RMSE, pooled across all three problems. Holm correction is across the full family of 24 tests (4 MOGPs $\times$ 2 objectives $\times$ 3 n-values).

Table 7: RMSE Wilcoxon signed-rank against the independent-GP baseline. Each row pools over WOSGZ1/2/3 and seeds (30 paired differences); rows are reported per n. Holm correction across the full 24-test family. Win count is out of 10 seeds $\times$3 problems = 30 matchups per (model, n, objective).

n	Model	Obj	p_{raw}	p_{Holm}	Wins/30
60	GPyTorch ICM (R=1)	f_1	<0.001	0.008**	22
60	GPyTorch ICM (R=1)	f_2	<0.001	<0.001***	27
60	GPyTorch LMC (R=2)	f_1	0.024	0.378	22
60	GPyTorch LMC (R=2)	f_2	0.030	0.394	19
60	BoTorch MultiTaskGP	f_1	<0.001	0.004**	22
60	BoTorch MultiTaskGP	f_2	<0.001	0.003**	23
60	MOGPTK SM-LMC	f_1	1.000	1.000	0
60	MOGPTK SM-LMC	f_2	1.000	1.000	0
120	GPyTorch ICM (R=1)	f_1	0.002	0.041*	24
120	GPyTorch ICM (R=1)	f_2	0.002	0.032*	21
120	GPyTorch LMC (R=2)	f_1	0.127	1.000	18
120	GPyTorch LMC (R=2)	f_2	0.175	1.000	19
120	BoTorch MultiTaskGP	f_1	0.028	0.392	21
120	BoTorch MultiTaskGP	f_2	0.026	0.392	19
120	MOGPTK SM-LMC	f_1	1.000	1.000	0
120	MOGPTK SM-LMC	f_2	1.000	1.000	0
230	GPyTorch ICM (R=1)	f_1	0.749	1.000	15
230	GPyTorch ICM (R=1)	f_2	0.901	1.000	15
230	GPyTorch LMC (R=2)	f_1	0.006	0.108	24
230	GPyTorch LMC (R=2)	f_2	0.015	0.249	24
230	BoTorch MultiTaskGP	f_1	0.948	1.000	11
230	BoTorch MultiTaskGP	f_2	0.742	1.000	17
230	MOGPTK SM-LMC	f_1	1.000	1.000	0
230	MOGPTK SM-LMC	f_2	1.000	1.000	0

RMSE tells a different story from NLPD. ICM and BoTorch MTGP win Holm-significantly at $n = 60$; ICM still wins at $n = 120$ while MTGP slips to uncorrected-only significance; and at

$n = 230$ the three surviving MOGPs are statistically tied with the baseline under Holm correction. ICM's and MTGP's ratios drift slightly above 1 (1.02 and 1.03 for ICM, 1.08 and 1.07 for MTGP), while LMC retains the only sub-1 ratios (0.976 and 0.973) but its advantage fails Holm correction. Honest reading of the $n = 230$ row: ICM trades a small RMSE penalty against the baseline ($\approx$ 2-3% above) for the calibration gains documented below; this is consistent with the standard pattern in which multi-output GPs add value most when data is scarce and the independent baseline catches up on mean accuracy as n grows, but it should be acknowledged as a trade-off rather than as "calibration gains at no mean-accuracy cost".

7.2.3 RMSE per-problem breakdown

Table 8 shows that the pooled RMSE-f_2 signal is driven disproportionately by WOSGZ1: all three non-SM-LMC MOGPs are Holm-significant on WOSGZ1, none are on WOSGZ2 or WOSGZ3. In contrast, the NLPD-f_2 advantage (Table 5) reaches Holm-significance on all three problems for ICM rank 1. This consistency of the NLPD advantage across the WOSGZ1-3 family, contrasted with the WOSGZ1-only RMSE advantage, is one of the reasons selection is grounded in NLPD rather than RMSE.

Table 8: Per-problem Wilcoxon signed-rank on RMSE-f_2 (pooling n-values by seed). Holm correction across 12 tests.

Problem	Model	p_{raw}	p_{Holm}	Win rate	Med. diff.
WOSGZ1	GPyTorch ICM (R=1)	0.001	0.012*	24/30	−0.014
WOSGZ1	GPyTorch LMC (R=2)	<0.001	0.003**	23/30	−0.010
WOSGZ1	BoTorch MultiTaskGP	0.002	0.015*	24/30	−0.014
WOSGZ1	MOGPTK SM-LMC	1.000	1.000	0/30	+0.332
WOSGZ2	GPyTorch ICM (R=1)	0.037	0.330	21/30	−0.009
WOSGZ2	GPyTorch LMC (R=2)	0.306	1.000	18/30	−0.004
WOSGZ2	BoTorch MultiTaskGP	0.145	0.723	17/30	−0.006
WOSGZ2	MOGPTK SM-LMC	1.000	1.000	0/30	+0.543
WOSGZ3	GPyTorch ICM (R=1)	0.085	0.597	18/30	−0.012
WOSGZ3	GPyTorch LMC (R=2)	0.046	0.368	21/30	−0.013
WOSGZ3	BoTorch MultiTaskGP	0.114	0.686	18/30	−0.020
WOSGZ3	MOGPTK SM-LMC	1.000	1.000	0/30	+0.515

7.2.4 Calibration

Table 9 reports coverage$_{95}$ (the joint-calibration metric used by the selection) and MACE-ρ (a diagnostic). All models, including the baseline, undercover the 95% credible ellipse, indicating joint predictive miscalibration consistent with overconfident posterior uncertainty and/or residual mean error: a limitation common to vanilla GPs, not specific to the setup used here. The MOGPs cover better than the baseline at every n, and the gap narrows but does not close as n grows: at $n = 230$ the baseline reaches 0.78 and ICM 0.84. Since coverage$_{95}$ uses the full $\mathbf{\Sigma}^{-1}$ via the Mahalanobis distance, this is the joint witness the marginal-NLPD metric does not provide. SM-LMC is the

clear outlier (coverage 0.22 to 0.37, severe overconfidence around a spurious joint structure) and is eliminated.

The MACE-ρ column is included as a diagnostic but is not used in the selection, for two reasons. First, the baseline value and the MOGP values are different quantities measured on the same scale: the baseline reports $|\rho_{\text{emp}}|$ over all 200 test points (a "missed-signal" indicator), while the MOGPs report a quintile-binned mean absolute calibration error of their predicted $\rho(\boldsymbol{x}^*)$ (a "miscalibration" indicator). Reading the column row-wise as "smaller is better, regardless of model type" would suggest the baseline is more cross-correlation calibrated than the MOGPs, which is misleading. Second, with $\sim$40 test points per quintile, the per-bin empirical Pearson correlations carry standard error $\approx 1/\sqrt{40} \approx 0.16$, so MACE-$\rho$ values below roughly 0.15 to 0.20 are within the resolution of the diagnostic. The column is retained because it is informative about SM-LMC's joint failure and about the qualitative direction of the MOGP rankings, but the joint-calibration weight in the selection argument sits on coverage_{95}.

Table 9: Calibration summary: each row reports the mean (std) across 10 seeds pooled over WOSGZ1/2/3. coverage_{95} (target 0.95) is the joint-calibration metric used in the selection. $|\rho_{\text{emp}}|$ is reported only for the independent baseline (the cross-objective structure it does not attempt to capture); MACE-ρ is reported only for MOGPs (calibration error of the predicted $\rho(\boldsymbol{x}^*)$, diagnostic only). See §7.1.2 for the noise-floor caveat on MACE-ρ.

n	Model	$\lvert\rho_{\text{emp}}\rvert$ (baseline)	MACE-ρ (MOGP, diag.)	coverage_{95}
60	Independent GP	0.108 (0.081)	—	0.596 (0.068)
60	GPyTorch ICM (R=1)	—	0.274 (0.106)	0.736 (0.042)
60	GPyTorch LMC (R=2)	—	0.256 (0.075)	0.635 (0.047)
60	BoTorch MultiTaskGP	—	0.219 (0.075)	0.755 (0.050)
60	MOGPTK SM-LMC	—	0.418 (0.170)	0.370 (0.255)
120	Independent GP	0.090 (0.068)	—	0.702 (0.037)
120	GPyTorch ICM (R=1)	—	0.224 (0.086)	0.793 (0.039)
120	GPyTorch LMC (R=2)	—	0.239 (0.072)	0.725 (0.038)
120	BoTorch MultiTaskGP	—	0.192 (0.062)	0.775 (0.041)
120	MOGPTK SM-LMC	—	0.372 (0.165)	0.324 (0.210)
230	Independent GP	0.086 (0.063)	—	0.783 (0.023)
230	GPyTorch ICM (R=1)	—	0.198 (0.077)	0.836 (0.027)
230	GPyTorch LMC (R=2)	—	0.223 (0.072)	0.799 (0.030)
230	BoTorch MultiTaskGP	—	0.232 (0.113)	0.803 (0.056)
230	MOGPTK SM-LMC	—	0.324 (0.154)	0.222 (0.135)

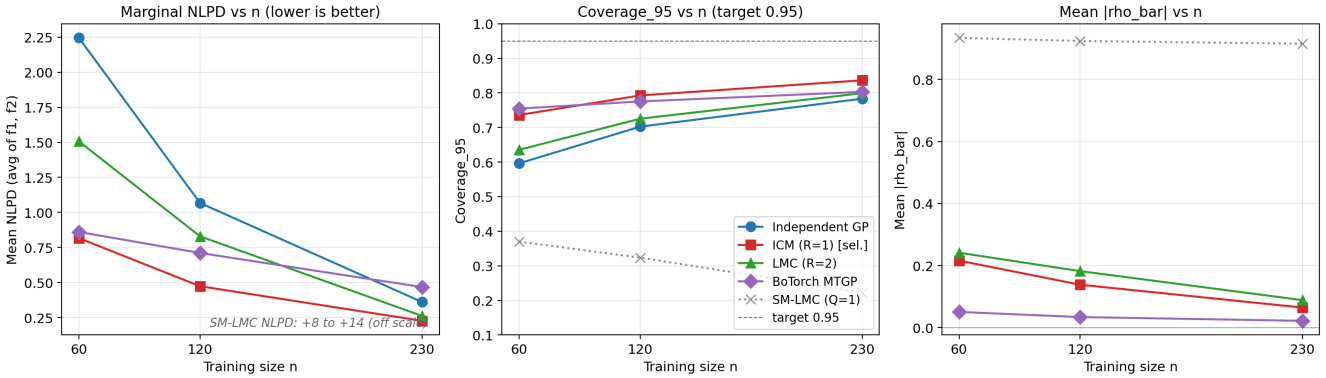

Figure 6: Per-model summary across $n \in \{60, 120, 230\}$, averaged over 10 seeds and WOSGZ1-3. *Left:* marginal NLPD averaged over (f_1, f_2) (co-primary; lower is better). *Center:* coverage$_{95}$ (co-primary; target 0.95, dashed). *Right:* mean $|\bar{\rho}|$ across test points (diagnostic; the independent baseline is identically zero by construction). SM-LMC is plotted with a dotted line because its values are consistently outside the viable range; it is included to show the contrast. ICM rank 1 (red, square markers) is the only candidate that remains co-primary-Pareto-non-dominated at $n = 120$ (descriptive) and $n = 230$ (pairwise-tested in Table 11); at $n = 60$ MTGP edges ICM by 0.02 on coverage but ICM beats MTGP on NLPD-f_1, so neither strictly dominates the other at the smallest n.

7.2.5 Pairwise MOGP comparisons at $n = 230$

Table 10 reports two-sided pairwise Wilcoxon tests among the four MOGPs at $n = 230$ (the regime most relevant to the BO loop), pooled across problems. Holm correction is across the $\binom{4}{2} \times 2 = 12$ tests.

Table 10: Two-sided pairwise Wilcoxon on RMSE at $n = 230$, pooled across WOSGZ1/2/3 by seed. Positive median difference means the row model has larger (worse) RMSE than the column model. Holm correction across 12 tests.

Row (a)	Col (b)	Obj	Med. $(a-b)$	p_{raw}	p_{Holm}
ICM (R=1)	LMC (R=2)	f_1	+0.010	0.038	0.154
ICM (R=1)	LMC (R=2)	f_2	+0.016	0.007	0.043*
ICM (R=1)	MTGP	f_1	−0.002	0.213	0.426
ICM (R=1)	MTGP	f_2	+0.001	0.730	0.730
ICM (R=1)	SM-LMC	f_1	−0.568	<0.001	<0.001***
ICM (R=1)	SM-LMC	f_2	−0.550	<0.001	<0.001***
LMC (R=2)	MTGP	f_1	−0.010	0.009	0.046*
LMC (R=2)	MTGP	f_2	−0.011	0.080	0.241
LMC (R=2)	SM-LMC	f_1	−0.559	<0.001	<0.001***
LMC (R=2)	SM-LMC	f_2	−0.577	<0.001	<0.001***
MTGP	SM-LMC	f_1	−0.538	<0.001	<0.001***
MTGP	SM-LMC	f_2	−0.557	<0.001	<0.001***

At $n = 230$, ICM, LMC, and MTGP are broadly indistinguishable on RMSE: the only Holm-significant pairwise-RMSE differences among the three are LMC marginally better than ICM on f_2 and LMC marginally better than MTGP on f_1. All three are clearly better than SM-LMC. The RMSE ranking among viable MOGPs at $n = 230$ is therefore close to a three-way tie; the discriminating evidence for the correlated-acquisition use case is reported in Table 11, which adds two-sided pairwise Wilcoxon tests at $n = 230$ on the co-primary metrics (marginal NLPD averaged over f_1, f_2, and coverage$_{95}$), pooled across WOSGZ1-3 by seed.

Table 11: Two-sided pairwise Wilcoxon at $n = 230$ on the co-primary metrics (marginal NLPD, averaged over f_1 and f_2; coverage$_{95}$), pooled across WOSGZ1-3 by seed. Holm correction across the three tests within each metric. Median is of the row-minus-column difference. For NLPD, negative values favor the row; for coverage$_{95}$, positive values favor the row.

Row (a)	Col (b)	Metric	Med. $(a-b)$	p_{raw}	p_{Holm}
ICM (R=1)	LMC (R=2)	NLPD	+0.003	0.360	0.360
ICM (R=1)	MTGP	NLPD	−0.215	<0.001	<0.001***
LMC (R=2)	MTGP	NLPD	−0.162	<0.001	<0.001***
ICM (R=1)	LMC (R=2)	cov$_{95}$	+0.030	<0.001	<0.001***
ICM (R=1)	MTGP	cov$_{95}$	+0.035	0.001	0.002**
LMC (R=2)	MTGP	cov$_{95}$	+0.000	0.845	0.845

This is the discriminator the chapter rests on at $n = 230$. ICM and LMC are statistically tied on marginal NLPD ($p_{\text{Holm}} = 0.36$, median difference +0.003, win rate 15/30) but ICM has significantly higher coverage$_{95}$ than LMC ($p_{\text{Holm}} < 0.001$, median +0.030, win rate 26/30). ICM beats MTGP on both: NLPD ($p_{\text{Holm}} < 0.001$, median −0.215, win rate 28/30) and coverage$_{95}$ ($p_{\text{Holm}} = 0.002$, median +0.035, win rate 23/30). Under the co-primary pair (marginal NLPD, coverage$_{95}$), ICM weakly dominates LMC (tied on NLPD, wins on coverage), LMC weakly dominates MTGP (wins NLPD, tied on coverage), and ICM strictly dominates MTGP (wins both). The cross-MOGP ordering at $n = 230$ is therefore ICM $\succeq$ LMC $\succeq$ MTGP, with ICM the unique co-primary-Pareto-non-dominated MOGP and the LMC-vs-MTGP tie-breaker not load-bearing for the selection.

What the pairwise tests do not establish. These tests use marginal NLPD rather than the joint bivariate NLPD that would directly correspond to the acquisition's input. Per-test-point posterior predictions $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ were not persisted by the benchmark script, so a numerical joint-NLPD computation would require re-fitting the models; this is left for a follow-up. The joint-marginal correction terms of §7.1.2 are bounded by the observed $|\bar{\rho}|$ range and aligned in sign by the volume term, and coverage$_{95}$ provides a direct $\boldsymbol{\Sigma}^{-1}$-using joint-calibration witness, but the chapter does not claim a tested joint-NLPD ranking. The selection therefore rests on co-primary marginal NLPD + coverage$_{95}$, with joint NLPD covered analytically by the decomposition argument rather than tested numerically.

7.2.6 Descriptive reference

Table 12 aggregates all metrics as means across 10 seeds and 3 problems, per n. GPyTorch ICM rank 1 (the selected MOGP; see Section 7.4) is highlighted in **bold**.

Table 12: Descriptive summary: means across 10 seeds, aggregated over WOSGZ1/2/3, per n. Ratio is MOGP RMSE over baseline RMSE. Bold rows mark GPyTorch ICM rank 1, the selected MOGP, and are not a column-wise "best" indicator: LMC has lower RMSE than ICM at $n = 230$, for example, but ICM is selected on the co-primary metrics. See §7.4.

n	Model	Ratio (f_1, f_2)	NLPD (f_1, f_2)	Fit (s)	$\lvert\bar{\rho}\rvert$	R_{12}	MACE-ρ	cov_{95}
60	Independent GP	(1.000, 1.000)	(+2.19, +2.30)	8.0	—	—	0.108	0.596
60	**ICM (R=1)**	**(0.945, 0.934)**	**(+0.81, +0.82)**	**19.9**	**0.216**	**+0.231**	**0.274**	**0.736**
60	LMC (R=2)	(0.976, 0.966)	(+1.50, +1.51)	63.6	0.241	+0.078	0.256	0.635
60	BoTorch MultiTaskGP	(0.946, 0.937)	(+0.90, +0.82)	12.5	0.050	+0.218	0.219	0.755
60	MOGPTK SM-LMC	(1.993, 2.033)	(+8.16, +9.01)	33.5	0.934	—	0.418	0.370
120	Independent GP	(1.000, 1.000)	(+1.09, +1.05)	35.3	—	—	0.090	0.702
120	**ICM (R=1)**	**(0.959, 0.953)**	**(+0.50, +0.44)**	**30.0**	**0.138**	**+0.086**	**0.224**	**0.793**
120	LMC (R=2)	(0.983, 0.991)	(+0.84, +0.82)	100.7	0.182	+0.035	0.239	0.725
120	BoTorch MultiTaskGP	(0.977, 0.964)	(+0.70, +0.72)	26.6	0.034	+0.108	0.192	0.775
120	MOGPTK SM-LMC	(2.305, 2.295)	(+14.3, +14.2)	91.5	0.924	—	0.372	0.324
230	Independent GP	(1.000, 1.000)	(+0.38, +0.34)	256.3	—	—	0.086	0.783
230	**ICM (R=1)**	**(1.019, 1.031)**	**(+0.23, +0.22)**	**83.4**	**0.065**	**+0.043**	**0.198**	**0.836**
230	LMC (R=2)	(0.976, 0.973)	(+0.28, +0.24)	294.5	0.089	+0.044	0.223	0.799
230	BoTorch MultiTaskGP	(1.078, 1.071)	(+0.47, +0.46)	59.0	0.022	+0.098	0.232	0.803
230	MOGPTK SM-LMC	(2.928, 3.017)	(+11.7, +10.9)	309.9	0.915	—	0.324	0.222

7.3 Analysis

RMSE is non-monotonic in n; ICM ties the baseline at $n = 230$. At $n \in \{60, 120\}$ the MOGPs win on RMSE under Holm correction (ICM on both n; MTGP only at $n = 60$); at $n = 230$ the three viable MOGPs are statistically tied with the baseline, and ICM's RMSE ratios drift slightly above 1 (1.02, 1.03). It is a standard pattern that multi-output GPs adding value most when data is scarce, with the independent baseline catching up on mean accuracy as n grows. The honest reading at the BO-relevant training size is that ICM trades a $\approx$ 2-3% RMSE penalty against the baseline for the calibration gains documented below; the calibration gain (coverage 0.836 vs 0.783) is the larger effect under an acquisition that consumes the full posterior, but RMSE-sensitive applications would weight the choice differently.

ICM rank 1 wins or ties on the co-primary pair at every training size. On marginal NLPD pooled across n, ICM rank 1 reaches $p_{\text{Holm}} < 0.001$ on every (problem, objective) cell, with win rates 27-29/30 and median NLPD reductions -0.25 to -0.61. On coverage_{95}, ICM is the best-covering MOGP at $n = 120$ and $n = 230$ (0.793 and 0.836); at $n = 60$ MTGP edges ICM on coverage by 0.02 (0.755 vs 0.736), but at the same n ICM beats MTGP on NLPD-f_1 and ties on NLPD-f_2. Strict component-wise Pareto dominance therefore does not hold at $n = 60$ between ICM and MTGP in either direction; at $n \in \{120, 230\}$ ICM wins or ties both co-primary metrics against every viable alternative, and the pairwise tests at $n = 230$ (Table 11) make this concrete: ICM

strictly dominates MTGP (wins NLPD and coverage), and weakly dominates LMC (ties NLPD, wins coverage). No alternative MOGP matches this pattern across the full n-grid; MTGP's small coverage edge at $n = 60$ is the single co-primary metric on which an alternative beats ICM, and it disappears at the larger BO-relevant training sizes.

WOSGZ1 drives the pooled RMSE signal; NLPD generalizes across the WOSGZ1-3 sample. On RMSE-f_2, all three non-SM-LMC MOGPs are Holm-significant only on WOSGZ1; on NLPD-f_2 ICM rank 1 is Holm-significant on every WOSGZ1 to WOSGZ3 cell. Selection on a metric that generalizes across the benchmark sample is more robust than selection on a metric driven by one of the three problems.

The benchmark is a size-matched proxy, not a distribution-matched proxy, for the BO regime. All training data in this chapter is drawn by Latin Hypercube sampling. The BO loop of Section 8 produces acquisition-driven training sets that concentrate near the Pareto front rather than spreading uniformly over the input space. The ICM posterior fitted on acquisition-driven data therefore differs in distribution from the LHS-fitted posterior characterized here; in particular, the per-point $|\rho(\boldsymbol{x})|$ in the BO loop is expected to differ from the LHS-test $|\bar{\rho}|$ values reported in Table 12 (mean 0.065 at $n = 230$). The mean $|\bar{\rho}|$ on uniform LHS test points is not a sufficient statistic for the cross-objective signal the acquisition optimizer sees in the BO inner loop, where evaluations concentrate in regions where the posterior carries informative cross-objective structure. Characterizing the BO-loop $|\bar{\rho}|$ directly would require running the BO loop, which is the experiment of Chapter 8. The BO results in that chapter therefore both validate and contextualize the surrogate selection: ICM's wins on WOSGZ1 to WOSGZ8 demonstrate that the LHS benchmark's surrogate ranking transfers, but the magnitude of the gain in the BO loop reflects acquisition-driven distributions that the LHS benchmark cannot directly characterize.

Selection scope: WOSGZ1 to WOSGZ3 only. The surrogate is selected on WOSGZ1 to WOSGZ3 and then deployed on the full WOSGZ1 to WOSGZ8 family in Chapter 8. WOSGZ1 to WOSGZ3 differ in the parameter tuple (A, B, C, D, E) and the shared shape term but share the same bias-and-distance functional form as WOSGZ4 to WOSGZ8, so a surrogate that captures the WOSGZ1 to WOSGZ3 structure is plausible a priori for the rest of the family. What this benchmark cannot rule out is that one of WOSGZ4 to WOSGZ8 has qualitatively different posterior cross-objective structure that might favor a different MOGP; the BO comparison in Chapter 8 is the empirical answer to that question, and the eight-problem WOSGZ wins reported there are consistent with the WOSGZ1 to WOSGZ3 selection generalizing.

BoTorch MTGP degrades faster than ICM as n grows. Pooled across problems, MTGP is Holm-significant on RMSE at $n = 60$ on both objectives but fades to uncorrected-only significance by $n = 120$ and to indistinguishable-from-baseline at $n = 230$ (ratios $1.08, 1.07$, win rates $11/30$ and $17/30$). The descriptive table shows the matching pattern on NLPD: MTGP's mean NLPD improvement over baseline shrinks from $(-1.29, -1.48)$ at $n = 60$ to $(-0.39, -0.33)$ at $n = 120$ to roughly $(+0.09, +0.12)$ at $n = 230$: by the largest training size, MTGP is no longer beating the baseline on NLPD either. MTGP is retained as a small-n ablation alternative but ICM is selected as the primary surrogate.

LMC is the only MOGP with RMSE ratio < 1 at $n = 230$. LMC's RMSE advantage at $n = 230$ (ratios $0.976, 0.973$) is the only sub-1 result in that regime, although it fails Holm correction ($p_{\text{Holm}} = 0.11, 0.25$). On NLPD, LMC reaches Holm-significance on WOSGZ1 (both objectives) and on f_1 at all problems, but is generally weaker than ICM and in some cells (WOSGZ3 NLPD-f_2) fails Holm. LMC is retained as an ablation alternative for downstream experiments.

SM-LMC ($Q = 1$) is empirically mismatched on this benchmark. On every (problem, n, metric) cell, the $Q = 1$ SM-LMC configuration tested here is worse than the baseline: RMSE ratios 1.99 to 3.02, median NLPD excesses +1.34 to +2.38 per objective, and coverage_{95} of 0.22 to 0.37. Its $|\bar{\rho}| \approx 0.93$ reflects spurious near-perfect posterior correlation around miscalibrated uncertainty. Spectral-mixture kernels target data with oscillatory or non-stationary structure, which the smooth bias-and-distance form of WOSGZ is not, but the conclusion here is about the empirical performance of this specific configuration on this benchmark, not about the inherent unsuitability of spectral-mixture LMC for MOBO surrogates. SM-LMC is eliminated for the present BO comparison.

The baseline catches up on mean accuracy but not on joint calibration. Coverage at $n = 230$ is 0.78 for the baseline vs. 0.84 for ICM: both well below the 0.95 target, but the relative gap persists rather than closing. Combined with ICM's marginal-NLPD advantage, this means the surrogate-quality gain a 170-iteration BO loop running with $n \approx 230$ training points sees from an MOGP shows up most in joint posterior calibration rather than in point predictions. The correlated acquisition uses that joint posterior directly. A calibration gain when mean accuracy is statistically tied is exactly the regime in which switching surrogates should matter for the acquisition.

7.4 Selection

Primary selection: GPyTorch ICM rank 1. The correlated ε-PoHVI acquisition function consumes the full bivariate predictive $\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x}))$, so surrogate quality is measured by the co-primary pair (marginal NLPD, coverage_{95}) read together via the joint-marginal NLPD decomposition of §7.1.2. Under marginal NLPD pooled across n, ICM rank 1 reaches $p_{\text{Holm}} < 0.001$ on every (problem, objective) cell, with median reductions -0.25 to -0.61 and win rates 27-29/30 (Tables 5-6). Under coverage_{95}, ICM is the best-covering MOGP at $n = 120$ and $n = 230$ (0.793 and 0.836, vs. baseline 0.702 and 0.783). Under RMSE, ICM matches the baseline at $n = 230$ (ratios $1.02, 1.03$, statistically tied under Holm) and outperforms it at $n \in \{60, 120\}$. ICM is the only MOGP that passes both co-primary checks at all three training sizes.

The BO experiment in Section 8 pairs corr-ε-PoHVI with GPyTorch ICM rank 1; the ε-PoHVI baseline keeps its independent-GP surrogate.

8 Bayesian Optimization Comparison

The verification chapter (Section 6.3) showed that the correlated ε-PoHVI distribution is computed correctly, and the surrogate chapter (Section 7) selected GPyTorch ICM rank 1 as the multi-output GP that supplies the posterior covariance to that distribution. The final empirical question is whether the resulting acquisition function actually improves Bayesian-optimization performance over the **ε-PoHVI** baseline of Wang et al. [27]. This chapter answers that question on the full test suite of Section 6.2, five ZDT problems and eight WOSGZ problems, with 15 seeds per (problem, method).

8.1 Setup

8.1.1 Algorithms

- **ε-PoHVI baseline**. Two independent `sklearn GaussianProcessRegressor` surrogates (one per objective) with Matérn-$\frac{3}{2}$ kernel, $5d = 150$ L-BFGS-B restarts, paired with the independent HVI distribution of Wang et al. [27].
- **corr-ε-PoHVI (this thesis)**. GPyTorch ICM rank 1 surrogate (Section 7.4) with Matérn-$\frac{3}{2}$ base kernel, 1000 Adam iterations, 3 random restarts, paired with the correlated HVI distribution of Chapter 3.

The ε-PoHVI baseline is inherited unmodified from the authors' released repository (see Section 6.1 for provenance details). The ε schedule $\varepsilon_t = \varepsilon_0 \exp(-c\,t)$ with $\varepsilon_0 = 0.05$ and $c = 0.02$ is the ε-PoHVI-scaling schedule of Wang et al. [27], §4; the counter t is the BO iteration index, starting at $t = 0$ immediately after the LHS initialization. The schedule is identical for both compared methods. The BO comparison therefore isolates the effect of replacing the independent-GP / independent-HVI pipeline with the multi-output-GP / correlated-HVI pipeline, while holding the ε_t schedule and the acquisition optimizer fixed.

8.1.2 Protocol

- **Test problems.** All 13 environments listed in Section 6.2 are used with equal weight: **ZDT1, ZDT2, ZDT3, ZDT4, ZDT6** and **WOSGZ1-WOSGZ8**, each at $n_{\text{var}} = 30$.
- **BO budget.** $n_{\text{init}} = 60$ Latin Hypercube initial samples followed by $n_{\text{iter}} = 170$ batch-1 BO iterations, for a total of $n_{\text{evals}} = 230$ objective evaluations per seed.
- **Acquisition optimizer.** CMA-ES with population $\lambda = 7$, 2,000 generations per restart, 3 restarts, BIPOP enabled. Identical settings for both methods; only the surrogate-posterior input to the acquisition differs.
- **Seeds.** 15 paired seeds per (problem, method), indexed 0-14.

8.1.3 Reference points: two distinct roles

The hypervolume indicator requires a reference point $\boldsymbol{r}$, and this experiment uses two different choices for two different purposes.

Acquisition reference $\boldsymbol{r}_{\text{acq}}$. The ε-PoHVI acquisition function computes $\Pr\big[\Delta\text{HV}(\boldsymbol{x}) > \varepsilon_t \mid \mathcal{D}\big]$, where ΔHV is hypervolume improvement defined relative to $\boldsymbol{r}_{\text{acq}}$. The values used are listed in Table 13, taken directly from Wang et al.'s released `args.yml` archive: $\boldsymbol{r}_{\text{acq}} = [1.2, 1.2]$ for WOSGZ and $[15, 15]$ for ZDT. This is the reference the BO is optimizing towards; changing it would change algorithm behavior.

Plotting reference $\boldsymbol{r}_{\text{plot}}$. A second reference is used only for post-hoc visualization of convergence curves. At the tight $\boldsymbol{r}_{\text{acq}} = [1.2, 1.2]$, early WOSGZ LHS points often dominate no part of the reference box, so the hypervolume at those iterations can be zero, making $\log_{10}(\text{HV}_\star - \text{HV})$ curves uninformative at the start. ZDT4's f_2 values can reach into the hundreds, so a plotting ref of $[15, 15]$ would yield zero HV for most ZDT4 evaluations; $[0.5, 389.802]$ is used for ZDT4 only, taken from the plotting script in Wang et al.'s released repository [27]. $\boldsymbol{r}_{\text{plot}}$ values are used only for within-problem, paired method comparisons; no cross-problem comparison of absolute HV values is made from $\boldsymbol{r}_{\text{plot}}$.

Table 13: Per-problem reference points used in the BO experiment. $\boldsymbol{r}_{\text{acq}}$ is passed to the ε-PoHVI acquisition; $\boldsymbol{r}_{\text{plot}}$ is used only for post-hoc visualization of HV trajectories so that early LHS iterations contribute non-zero hypervolume.

Family	Problem	$\boldsymbol{r}_{\text{acq}}$	$\boldsymbol{r}_{\text{plot}}$
ZDT	ZDT1, ZDT2, ZDT3, ZDT6	$[15, 15]$	$[15, 15]$
ZDT	ZDT4	$[15, 15]$	$[0.5, 389.802]$
WOSGZ	WOSGZ1, WOSGZ2, WOSGZ6, WOSGZ8	$[1.2, 1.2]$	$[2.5, 2.5]$
WOSGZ	WOSGZ3	$[1.2, 1.2]$	$[1, 2]$
WOSGZ	WOSGZ4	$[1.2, 1.2]$	$[15, 4.5]$
WOSGZ	WOSGZ5	$[1.2, 1.2]$	$[8, 3]$
WOSGZ	WOSGZ7	$[1.2, 1.2]$	$[10, 2]$

8.1.4 Latin Hypercube lockdown

To make the paired seed-level comparison exact on initial conditions, the Latin Hypercube samples are pre-generated once per (problem, seed) combination and saved under `experiments/init_samples/<problem>/seed_SS.csv`. Both methods read the same CSV for seed s on problem p, eliminating LHS-draw variance as a confounder in the paired analysis and making the initial design reproducible across machines and across future ablation additions.

8.1.5 Statistical methodology

For each problem the per-seed final hypervolume at iteration 170 (evaluated at $\boldsymbol{r}_{\text{acq}}$) is computed and the paired Wilcoxon signed-rank test is applied to the 15 seed-level differences $d_s = \text{HV}^{(s)}_{\text{corr}} - \text{HV}^{(s)}_{\text{base}}$, testing both two-sided and the one-sided alternative H_1 : corr $>$ base. The exact Wilcoxon distribution is used (rather than its asymptotic normal approximation), which is feasible at $n = 15$ and removes the small-sample mis-calibration that the normal approximation can introduce. Across-problem multiple comparisons are corrected by the Holm-Bonferroni step-down procedure [18] on the family of 13 per-problem two-sided p-values. Significance markers throughout: * $p_{\text{Holm}} < 0.05$,

** $p_{\text{Holm}} < 0.01$, *** $p_{\text{Holm}} < 0.001$. As effect-size complements, each problem also reports the seed-level win/loss/tie count out of 15 paired comparisons and (in Table 15) the Hodges-Lehmann pseudo-median of $\{d_s\}$ with a Wilcoxon-based 95% CI; a low Holm-corrected p-value combined with a high win count and a CI that excludes zero is the strongest evidence pattern.

p-value floor at $n = 15$. At $n = 15$, when all 15 paired differences agree in sign, the exact two-sided Wilcoxon p-value saturates at its minimum attainable value $2/2^{15} = 6.1 \times 10^{-5}$. Several WOSGZ rows in Table 14 hit exactly this floor; their Holm-corrected p-values are therefore not separating those problems statistically: they are all "as significant as the test allows under $n = 15$ and an all-same-sign paired difference". The fact that 9 of the 13 problems pass $p_{\text{Holm}} < 0.001$ at this floor also means the WOSGZ wins are robust to many alternative correction strategies (any FWER- or FDR-controlling procedure with fewer than $\sim$100 tests in the family would still flag them).

Hyperparameter-optimizer disparity is conservative against the MOGP. The baseline gets $5d = 150$ L-BFGS-B restarts; the correlated method's ICM rank 1 surrogate gets 3 Adam restarts. This asymmetry is deliberately tilted in favor of the baseline: the baseline's optimizer sees 50× more restart budget than the MOGP's. Section 7.1 discusses the pilot studies behind these settings and the ICM-restart sensitivity check; the BO comparison here therefore reports a conservative MOGP win.

8.1.6 Baseline-comparison caveat

Absolute HV values in the runs reported here differ from those reported in Wang et al. [27] Table 4 for the WOSGZ family. The dominant contributor is a problem-definition divergence: the implementation used here follows Wang, Ong, Sun, Gupta & Zhang [28] (cross-verified against Santiago's independent implementation), whereas the released ICML-rebuttal code of Wang et al. [27] uses a variant with f_1 scaled differently from the 2019 definition; the per-point ratio (Wang et al. over the implementation used here) varies between 5× and 13×. These are not the same benchmark problem. A statistical comparison against Wang et al. Table 4 is therefore not possible for the WOSGZ family. The statistically controlled comparison reported here is paired-by-seed within a canonical implementation.

8.2 Results

8.2.1 Master final-HV summary

Table 14 reports the final hypervolume after 170 BO iterations across 15 seeds per (problem, method), evaluated at $\boldsymbol{r}_{\text{acq}}$, with the paired Wilcoxon p-value (two-sided) and Holm correction across the full 13-problem family. Sign of the gap is positive when corr-ε-PoHVI is ahead.

Table 14: Final HV at $\boldsymbol{r}_{\text{acq}}$ after 170 BO iterations, 15 seeds per cell. Mean gap = mean of paired differences (corr-ε-PoHVI − ε-PoHVI). W/L/T = win / loss / tie count of the 15 paired comparisons. p_{two}: raw paired Wilcoxon two-sided. p_{Holm}: Holm-Bonferroni correction across the 13-problem family.

Problem	corr-ε-PoHVI		ε-PoHVI		Mean	Med.	W/L/T	p_{two}	p_{Holm}
	mean	sd	mean	sd	gap	gap			
ZDT1	217.98	6.77	221.72	1.37	−3.74	−1.81	7/8/0	0.169	1.000
ZDT2	201.38	4.58	209.01	3.70	−7.63	−7.28	2/13/0	0.00085	0.0085**
ZDT3	228.53	4.55	225.82	6.58	+2.71	+3.60	11/4/0	0.0946	0.284
ZDT4†	0.00	0.00	0.00	0.00	0.00	0.00	0/0/15	1.000	1.000
ZDT6	216.29	3.63	141.28	11.11	+75.01	+76.71	15/0/0	6.1×10^{-5}	< 0.001***
WOSGZ1	0.666	0.021	0.520	0.068	+0.147	+0.148	15/0/0	6.1×10^{-5}	< 0.001***
WOSGZ2	0.635	0.029	0.441	0.065	+0.194	+0.196	15/0/0	6.1×10^{-5}	< 0.001***
WOSGZ3	0.545	0.038	0.365	0.061	+0.180	+0.188	15/0/0	6.1×10^{-5}	< 0.001***
WOSGZ4	0.510	0.028	0.279	0.045	+0.231	+0.241	15/0/0	6.1×10^{-5}	< 0.001***
WOSGZ5	0.470	0.009	0.257	0.052	+0.213	+0.211	15/0/0	6.1×10^{-5}	< 0.001***
WOSGZ6	0.463	0.007	0.232	0.036	+0.231	+0.237	15/0/0	6.1×10^{-5}	< 0.001***
WOSGZ7	0.314	0.034	0.204	0.053	+0.110	+0.089	15/0/0	6.1×10^{-5}	< 0.001***
WOSGZ8	0.967	0.038	0.718	0.077	+0.249	+0.271	15/0/0	6.1×10^{-5}	< 0.001***

† At $\boldsymbol{r}_{\text{acq}} = [15, 15]$ every ZDT4 seed produces zero HV for both methods (the optimizer never finds a point dominating the tight reference). The non-degenerate ZDT4 row at $\boldsymbol{r}_{\text{plot}} = [0.5, 389.802]$ gives (corr mean, base mean) = (14.80, 18.81), mean gap −4.00, W/L/T 7/7/1, $p_{\text{two}} = 0.50$, which is not Holm-significant; see the $\boldsymbol{r}_{\text{acq}}/\boldsymbol{r}_{\text{plot}}$ paragraph in §8.3. The remaining 12 rows are at $\boldsymbol{r}_{\text{acq}}$ as stated.

After Holm correction across the 13-problem family, corr-ε-PoHVI wins on every problem with exploitable inter-objective coupling: all eight WOSGZ problems plus ZDT6, all at $p_{\text{Holm}} < 0.001$. It is statistically indistinguishable on three of the remaining four ZDT problems (ZDT1, ZDT3, ZDT4); the single Holm-significant loss is on ZDT2 at $p_{\text{Holm}} = 0.0085$. Reading the result by family: **WOSGZ** 8/0/0 (wins / ties / losses out of 8 problems), **ZDT** 1/3/1 (out of 5). The empirical gain tracks the design distinction between the two suites, with WOSGZ-style coupling acting as a precondition for the correlated acquisition's advantage. Furthermore, the surrogate selection of Chapter 7 used WOSGZ1-3 only, so the wins on WOSGZ4-WOSGZ8 are independent validation that ICM rank 1 generalizes across the WOSGZ family rather than overfitting to the three problems used in selection (matching $|\bar{\rho}|_{\text{BO}}$ ranges in Table 15 reinforce this). Likewise, the hyperparameter-optimizer disparity in §8.1 ($5d = 150$ baseline restarts vs. 3 MOGP restarts) is tilted on purpose in favor of the baseline, so the BO comparison reports a conservative MOGP win.

Table 15: Relative gap (mean paired diff / baseline mean), Hodges-Lehmann pseudo-median of paired diffs with 95% Wilcoxon-based CI, iterations-to-target where "target" is 90% of the per-problem best mean HV either method achieves within the 170-iteration budget, and $|\bar{\rho}|$ along the actual BO trajectory (mean over BO iterations × seeds of the posterior $|\bar{\rho}|$ at the candidates the optimizer evaluated. "≥ 170" for iterations-to-target indicates the method does not reach 90% of best within the budget. ZDT4 is a degenerate case at the tight $\boldsymbol{r}_{\text{acq}} = [15, 15]$: many seeds yield zero HV for both methods and the 90%-of-best target collapses to 0; see the $\boldsymbol{r}_{\text{plot}}$ sensitivity paragraph in §8.3.

Problem	Rel. gap	H-L median	H-L 95% CI	It. to-90% (corr)	It. to-90% (base)	$\lvert\bar{\rho}\rvert_{\text{BO}}$
ZDT1	−1.7%	−4.15	[−6.76, +1.17]	1	1	0.003
ZDT2	−3.6%	−7.65	[−11.6, −3.74]	1	3	0.0003
ZDT3	+1.2%	+3.14	[−1.09, +6.66]	5	23	0.001
ZDT4	—	0.00	[0, 0]	—	—	0.014
ZDT6	+53%	+75.3	[+67.2, +83.0]	28	≥ 170	0.018
WOSGZ1	+28%	+0.149	[+0.109, +0.181]	110	≥ 170	0.168
WOSGZ2	+44%	+0.195	[+0.152, +0.237]	113	≥ 170	0.216
WOSGZ3	+49%	+0.180	[+0.159, +0.205]	102	≥ 170	0.184
WOSGZ4	+83%	+0.234	[+0.211, +0.256]	99	≥ 170	0.223
WOSGZ5	+83%	+0.213	[+0.182, +0.247]	71	≥ 170	0.178
WOSGZ6	+100%	+0.233	[+0.211, +0.252]	65	≥ 170	0.233
WOSGZ7	+54%	+0.097	[+0.071, +0.145]	112	≥ 170	0.116
WOSGZ8	+35%	+0.251	[+0.203, +0.301]	94	≥ 170	0.183

The relative-gap column reorders the magnitudes against the raw mean-gap column of Table 14: WOSGZ4-WOSGZ6 are the dominant wins in proportional terms (+83% to +100%), not ZDT6 (+53%) which dominates the raw mean-gap column simply because ZDT6's HV scale is two orders of magnitude larger than the WOSGZ scale. The Hodges-Lehmann CIs make the borderline ZDT results explicit: ZDT1 and ZDT3 straddle zero, ZDT2 is strictly negative, and the WOSGZ CIs are strictly positive with widths comparable to the WOSGZ standard deviations. The Hodges-Lehmann pseudo-medians differ slightly from the simple medians of paired differences in Table 14 (e.g. ZDT2 H-L −7.65 vs simple median −7.28); this is expected and reflects the difference between the two summary statistics. The iterations-to-target column reframes the WOSGZ wins as a data-efficiency story: corr-ε-PoHVI reaches 90% of the best mean HV either method achieves at iteration 65-113 on WOSGZ, while the baseline does not reach 90% of best on any WOSGZ problem within the 170-iteration budget. This metric is most informative on the WOSGZ family and ZDT6, where the budget is genuinely constraining; on ZDT1-3 the LHS initialisation already delivers most of the per-problem best HV and both methods reach the 90%-of-best target within the first ~ 25 iterations, so the column is reported but does not discriminate strongly there. Finally, the $|\bar{\rho}|_{\text{BO}}$ column (the mean posterior $|\bar{\rho}|$ at the BO candidates) gives a numerical witness for the WOSGZ-vs-ZDT split: WOSGZ trajectories produce $|\bar{\rho}|$ in the 0.12-0.23 range, ZDT1/ZDT2/ZDT3 are in the 10^{-4}-3×10^{-3} range (essentially zero), and ZDT4 and ZDT6 are intermediate at 0.014 and 0.018. Across the problems, the cross-objective signal is substantially larger on every WOSGZ problem (where the correlated method wins) than on the ZDT problems with no exploitable coupling (where the method ties or loses). The mapping is not perfect at the borderline: ZDT4 and ZDT6 carry comparable $|\bar{\rho}|_{\text{BO}}$ but produce very different outcomes (ZDT4 degenerate at $\boldsymbol{r}_{\text{acq}}$, ZDT6 a +53%

win), which says the cross-objective signal is a necessary condition but not a sufficient one. On ZDT4 the multimodal landscape and per-seed HV variance dominate the comparison even when the surrogate has some cross-objective signal to offer.

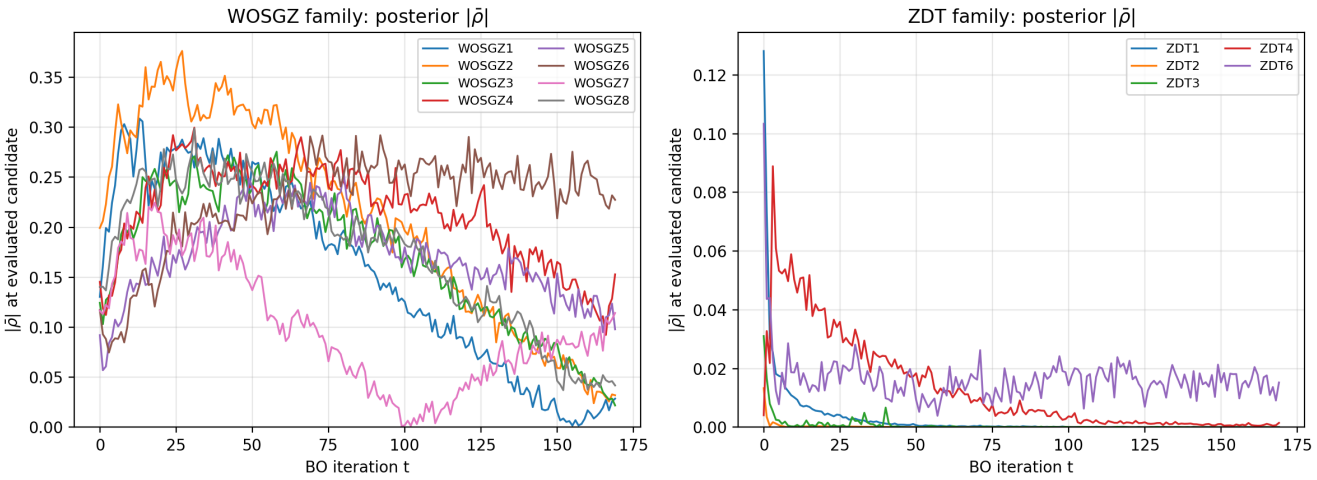

Figure 7: Posterior $|\bar{\rho}|$ along the BO trajectory, mean over 15 seeds of the per-iteration mean-of-candidate-points $|\bar{\rho}|$ logged in `corr_diagnostics.csv` for each seed. WOSGZ panels (left) show $|\bar{\rho}|$ in the 0.10 to 0.25 range; ZDT panels (right) show $|\bar{\rho}|$ near zero except ZDT4 and ZDT6 where it stays small but nonzero. This is the cross-objective signal that the correlated acquisition is built to exploit, and it is substantially larger than the LHS-test $|\bar{\rho}|$ values reported in Section 7.2 (e.g. 0.065 at $n = 230$ on uniform LHS test points, versus 0.16-0.23 on the BO trajectory): acquisition-driven training data concentrates in regions where the posterior carries informative cross-objective structure, exactly as Section 7.3 flagged.

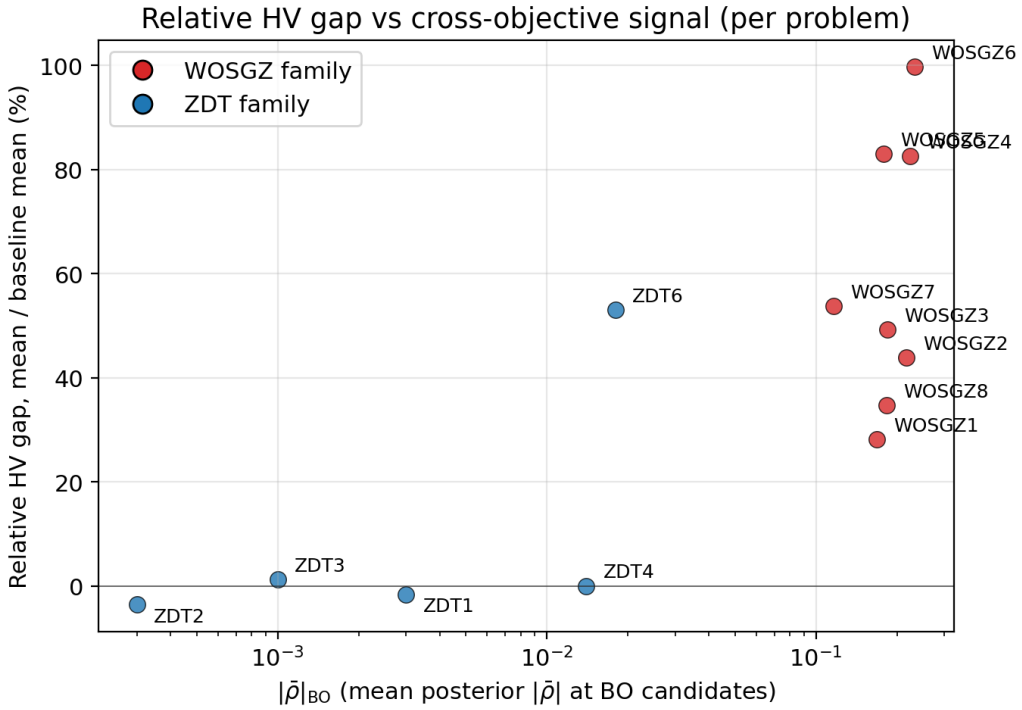

Figure 8: Relative HV gap (mean paired difference / baseline mean HV) versus the trajectory $|\bar{\rho}|_{\mathrm{BO}}$, one point per problem, log-scale on the horizontal axis. WOSGZ (red) and ZDT (blue) separate cleanly along $|\bar{\rho}|_{\mathrm{BO}}$, but the relationship is not monotone: ZDT4 and ZDT6 carry comparable $|\bar{\rho}|_{\mathrm{BO}}$ (0.014 vs 0.018) but very different relative gaps, indicating that cross-objective signal is a necessary but not sufficient condition for an HV win. The WOSGZ7 row sits below the rest of WOSGZ on $|\bar{\rho}|_{\mathrm{BO}}$ but still produces a +54% relative gain.

8.2.2 Per-problem convergence: WOSGZ family

Figure 9 reports the per-problem convergence of both methods on every WOSGZ problem. The y-axis is $\log_{10}(\mathrm{HV}^\star - \mathrm{HV}(t))$, where $\mathrm{HV}^\star$ is the best final hypervolume achieved by either method across all 15 seeds on that problem (a tight per-problem bound, not the analytical Pareto-front HV). Lower curves are better. The correlated method (blue) lies strictly below the baseline (orange) on every WOSGZ problem from the earliest iterations onward, with the gap opening within the first 25-50 iterations and persisting.

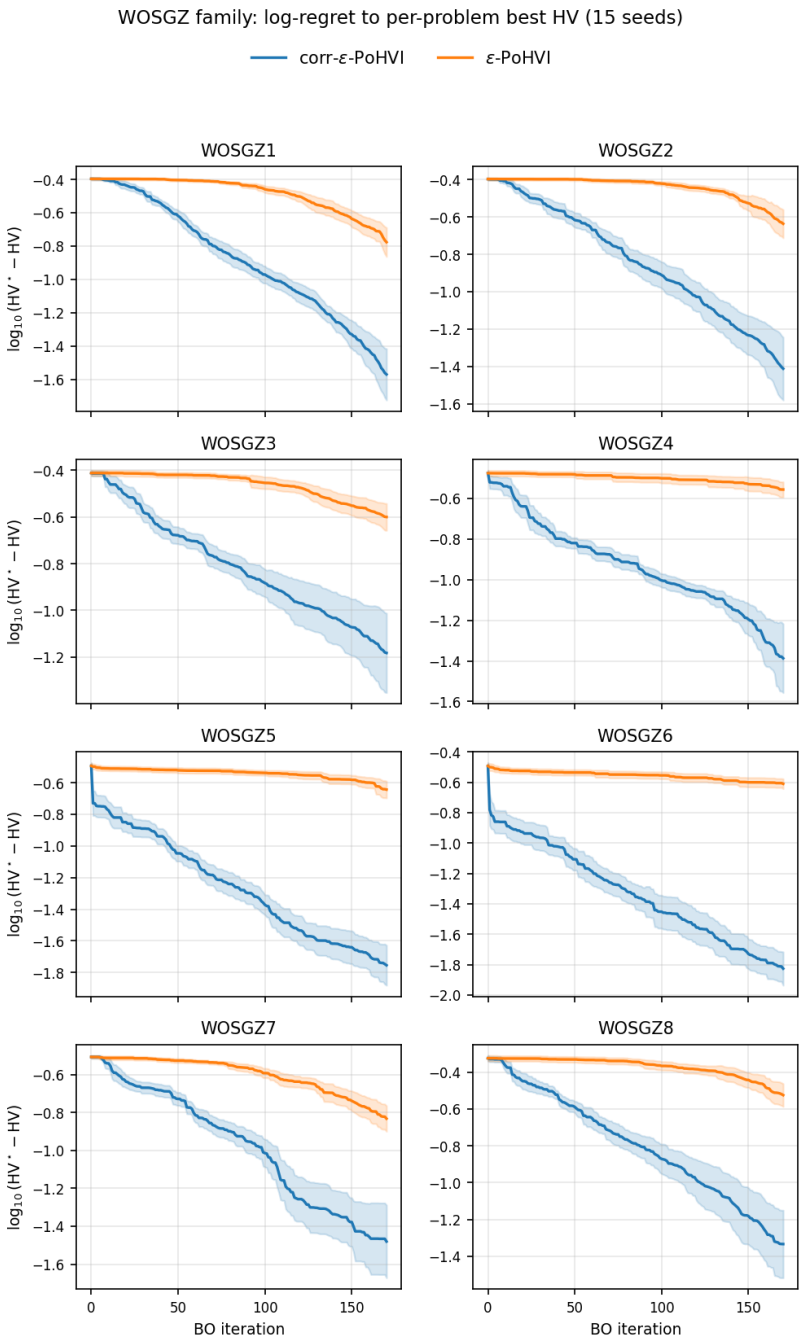

Figure 9: WOSGZ family convergence: $\log_{10}(\mathrm{HV}^\star - \mathrm{HV}(t))$ versus BO iteration, mean over 15 seeds with 95% CI ribbon. $\mathrm{HV}^\star$ is the per-problem best HV achieved by either method. Lower is better.

8.2.3 Per-problem convergence: ZDT family

Figure 10 shows the same convergence view for the ZDT family. The picture is mixed: corr-ε-PoHVI tracks the baseline closely on ZDT1, advances faster early then is overtaken on ZDT4, sits below the baseline throughout on ZDT3 and (decisively) ZDT6, and sits above the baseline on ZDT2.

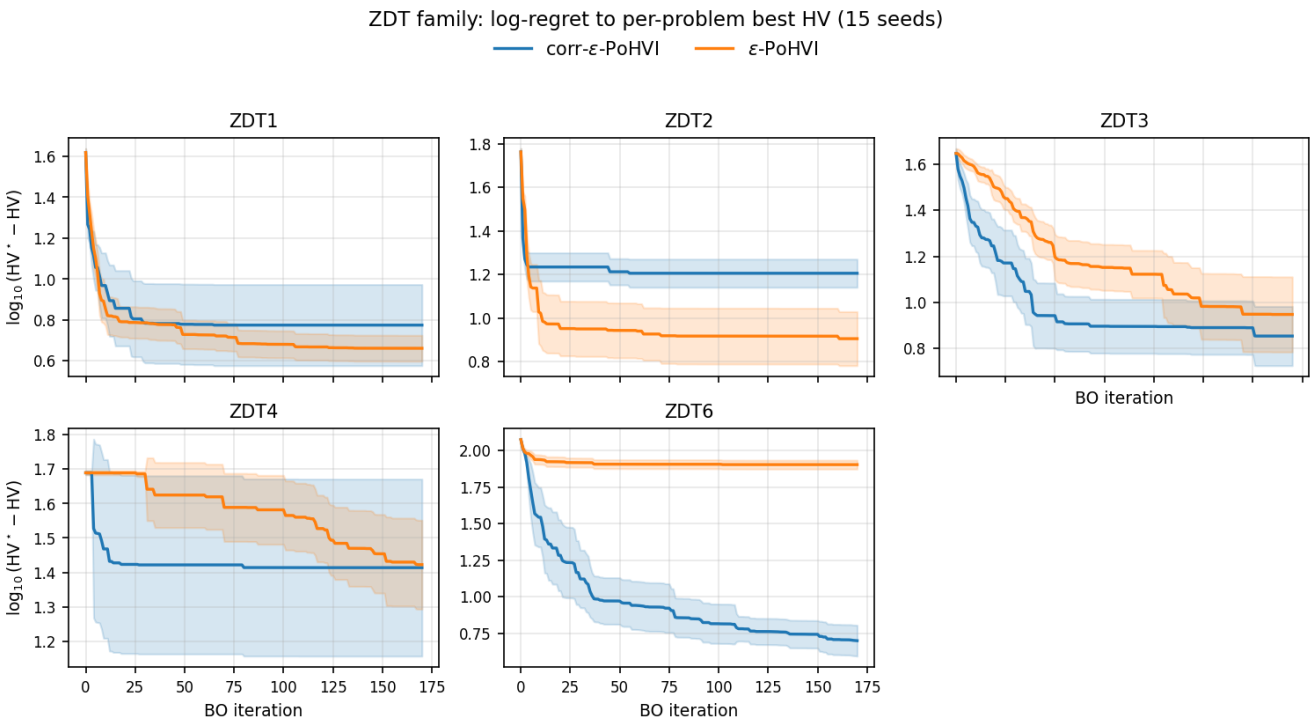

Figure 10: ZDT family convergence: $\log_{10}(\mathrm{HV}^\star - \mathrm{HV}(t))$ versus BO iteration, mean over 15 seeds with 95% CI ribbon. Same construction as Figure 9. Lower is better.

8.2.4 Paired HV gap

Figure 11 shows the per-iteration paired HV gap (corr-ε-PoHVI $-$ ε-PoHVI) for every problem in the suite. Positive values indicate the correlated method is ahead. The figure exposes the suite split at a glance: every WOSGZ panel shows a positive gap that grows during early iterations, and then either stabilises or modestly narrows in the late-iteration tail (WOSGZ1, WOSGZ7, and WOSGZ8 are visibly declining by iteration ~ 150, consistent with the baseline catching up on mean accuracy as n grows). ZDT panels are mostly flat near zero except ZDT6 (large positive gap throughout) and ZDT2 (persistently negative gap). The WOSGZ pattern is consistent with corr-ε-PoHVI buying faster convergence rather than a strictly better asymptote; the iterations-to-90%-target column of Table 15 quantifies this: corr reaches 90% of the best mean HV in 65 to 113 BO iterations on WOSGZ while the baseline does not reach the same target within the 170-iteration budget on any WOSGZ problem.

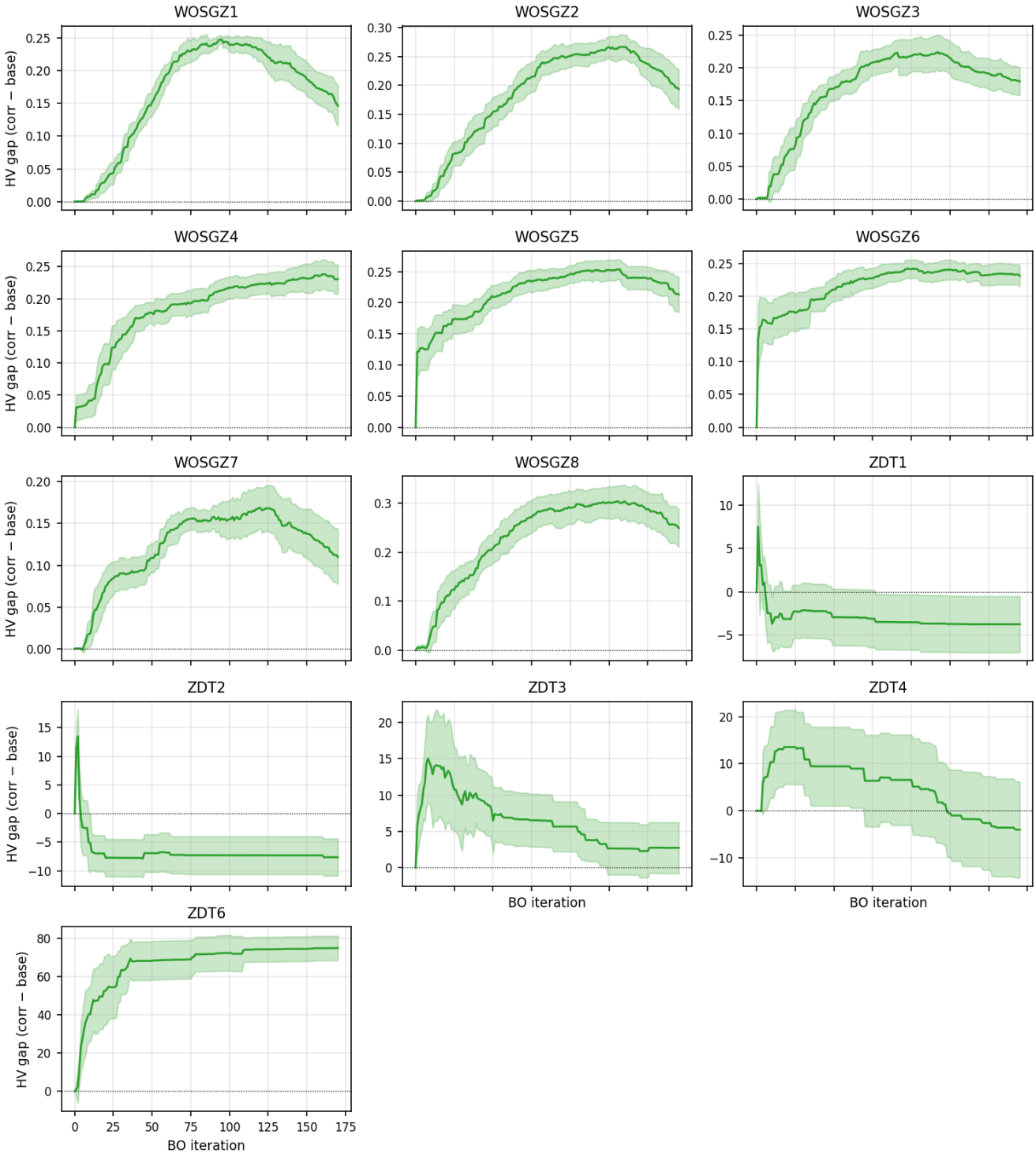

Figure 11: Paired HV gap per BO iteration (corr-ε-PoHVI − ε-PoHVI) on each problem, mean over 15 seeds with 95% CI ribbon. Positive = correlated method ahead; dotted line at zero.

8.3 Analysis

Suite-level split. The clearest finding is the suite split: every WOSGZ problem is a clean win for corr-ε-PoHVI (15/15 seeds, mean gap +0.11 to +0.25 HV units), whereas the ZDT family shows a mixed pattern. This matches the design intent of the two suites: WOSGZ is explicitly constructed to introduce inter-objective coupling through the shared shape term in Equation (142), while ZDT problems share dependence on x_1 but otherwise route the modeling burden through independent variable subsets, so the posterior cross-correlation $\rho(\boldsymbol{x})$ that the ICM surrogate tries to exploit is uninformatively small. ZDT's high raw-data correlation $\mathrm{Corr}(f_1, f_2) \approx -0.78$ across random samples reflects shared x_1 dependence rather than coupling that ICM can capture.

Magnitude grows with problem index in the WOSGZ family. The mean HV gap on WOSGZ ranges from +0.11 on WOSGZ7 to +0.25 on WOSGZ8, with the harder WOSGZ4-6 and WOSGZ8 in the +0.21-+0.25 band and the more constrained WOSGZ7 (denser high-frequency bias term) at the low end. In every case the absolute gap exceeds the per-method seed-level standard deviation, so the wins are not marginal.

Variance reduction. Across the WOSGZ family, the standard deviation of final HV is 2-8× smaller for corr-ε-PoHVI than for the baseline (e.g. WOSGZ5: 0.009 vs 0.052; WOSGZ6: 0.007 vs 0.036). On ZDT6 the same pattern holds: sd = 3.63 for corr-ε-PoHVI vs 11.11 for the baseline. For a practitioner running a single expensive BO campaign this reliability gain is arguably as consequential as the mean shift: the worst corr-ε-PoHVI seed on most WOSGZ problems beats the baseline's mean.

ZDT6 behaves like a WOSGZ problem. Among the ZDT family, only ZDT6 shows a strong win (+75 HV, 15/15 seeds, $p_{\mathrm{Holm}} < 0.001$). ZDT6 differs from the rest of ZDT in that its $g(\boldsymbol{x})$ has a non-trivial dependence on the entire $x_{2:d}$ vector through a sum of cosine modulations, introducing genuine cross-objective structure of the kind ICM can model. In other words, ZDT6 is closer to the WOSGZ regime than to the ZDT1/ZDT2 regime, and the result reflects that.

ZDT2 is a Holm-significant loss; ZDT1 borderline; ZDT4 a tie with a measurement caveat. ZDT2 is the only problem on which corr-ε-PoHVI loses with high confidence ($p_{\mathrm{Holm}} = 0.0085$, 13/15 seeds favor the baseline, mean gap −7.6 HV; Hodges-Lehmann −7.65 with 95% CI $[-11.6, -3.74]$). ZDT1 leans baseline (Hodges-Lehmann −4.15 with CI $[-6.76, +1.17]$) but the CI straddles zero and the test does not survive Holm correction. ZDT4 reports a degenerate 0/0/15 W/L/T split at $\boldsymbol{r}_{\mathrm{acq}} = [15, 15]$ because every seed of both methods fails to produce a point dominating the tight $\boldsymbol{r}_{\mathrm{acq}}$; the non-degenerate ZDT4 row at $\boldsymbol{r}_{\mathrm{plot}} = [0.5, 389.802]$ gives mean HVs 14.80 and 18.81 with 7/7/1 W/L/T and is reported in the footnote of Table 14.

Why ZDT2 in particular? ZDT1 and ZDT2 share the same lack of inter-objective coupling that ICM can capture (Figure 7 shows $|\bar{\rho}|_{\mathrm{BO}} \approx 0.003$ for ZDT1 and 0.0003 for ZDT2), yet ZDT1 is a tie and ZDT2 is a Holm-significant loss. The chapter does not have a confidently identified mechanism for the asymmetry, but three candidates are worth listing rather than papered over: (i) ZDT2's concave Pareto front may make the cell-partition geometry more sensitive to small

posterior misspecifications than ZDT1's convex front, so the same model-class capacity ICM brings is absorbed differently on the two problems; (ii) ICM at $n = 230$ carries a $\approx 3\%$ RMSE penalty against the baseline on f_2 (Section 7.2, Table 12), and ZDT2's f_2 governs the upper-left portion of the front, so the penalty is likely to bite harder there than on ZDT1; (iii) the fitted $|\bar{\rho}|$ on the BO trajectory for ZDT2 is on the order of 10^{-4}, but the sign of ρ at individual evaluated points may be systematically wrong, biasing the acquisition away from regions the baseline correctly explores. An ICM-marginal ablation (i.e. ICM means and variances but with $\rho \equiv 0$ substituted into the acquisition) would discriminate (ii) from (iii); that ablation was not run, so the mechanism attribution is left open.

Reference-point robustness: $\boldsymbol{r}_{\text{acq}}$ vs $\boldsymbol{r}_{\text{plot}}$. The Wilcoxon tests in Tables 14 and 15 use HV at $\boldsymbol{r}_{\text{acq}}$; the convergence figures plot HV at $\boldsymbol{r}_{\text{plot}}$ (§8.1.3). These can disagree, especially when $\boldsymbol{r}_{\text{acq}}$ is tight relative to the empirical f-range. Repeating the per-problem two-sided Wilcoxon at $\boldsymbol{r}_{\text{plot}}$ gives Holm-corrected p-values that match the $\boldsymbol{r}_{\text{acq}}$ table on every problem except ZDT4, where the $\boldsymbol{r}_{\text{acq}}$ result is degenerate ($p_{\text{Holm}} = 1$, 0/0/15 ties at $\boldsymbol{r}_{\text{acq}}$) and the $\boldsymbol{r}_{\text{plot}}$ result is non-degenerate but still not Holm-significant ($p_{\text{Holm}} = 0.50$). The visual ZDT4 lead in Figure 10 is at $\boldsymbol{r}_{\text{plot}}$ and does not translate to a final-HV advantage at either reference. On every other problem the two reference points produce the same qualitative ranking, so the headline conclusion is robust to the $\boldsymbol{r}_{\text{acq}}/\boldsymbol{r}_{\text{plot}}$ choice.

Computational overhead. Surrogate-fit cost (Table 12) is comparable: ICM at $n = 230$ averages 83.4 s vs the independent-GP baseline's 256.3 s, so the MOGP is actually faster than the baseline at large n because GPyTorch's Adam loop scales better than the baseline's $5d = 150$ L-BFGS-B restarts. The bivariate normal CDF inside the correlated acquisition is roughly 2 to 3$\times$ more expensive per call than the scalar normal CDF, but per-iteration acquisition optimization cost is dominated by CMA-ES candidate evaluations rather than by single CDF calls and is shared between methods. Per-iteration end-to-end BO wall-clock was not separately profiled, so the per-iteration total is reported as "comparable in these runs" rather than with a quantitative claim.

Per-seed structure of the wins. Figure 12 plots the per-seed $(\text{HV}_{\text{base}}, \text{HV}_{\text{corr}})$ pairs for every problem, separated by family. On the WOSGZ family every seed lies above the $y = x$ diagonal, the wins are uniform shifts across all 15 seeds, not driven by a small number of catastrophic baseline runs. On the ZDT family the points cluster on the diagonal except for ZDT6 (uniform shift, 15/15) and ZDT2 (uniform shift the other way, 13/15); ZDT4 is a stripe at low HV where many seeds yield near-zero HV at $\boldsymbol{r}_{\text{acq}}$ for both methods.

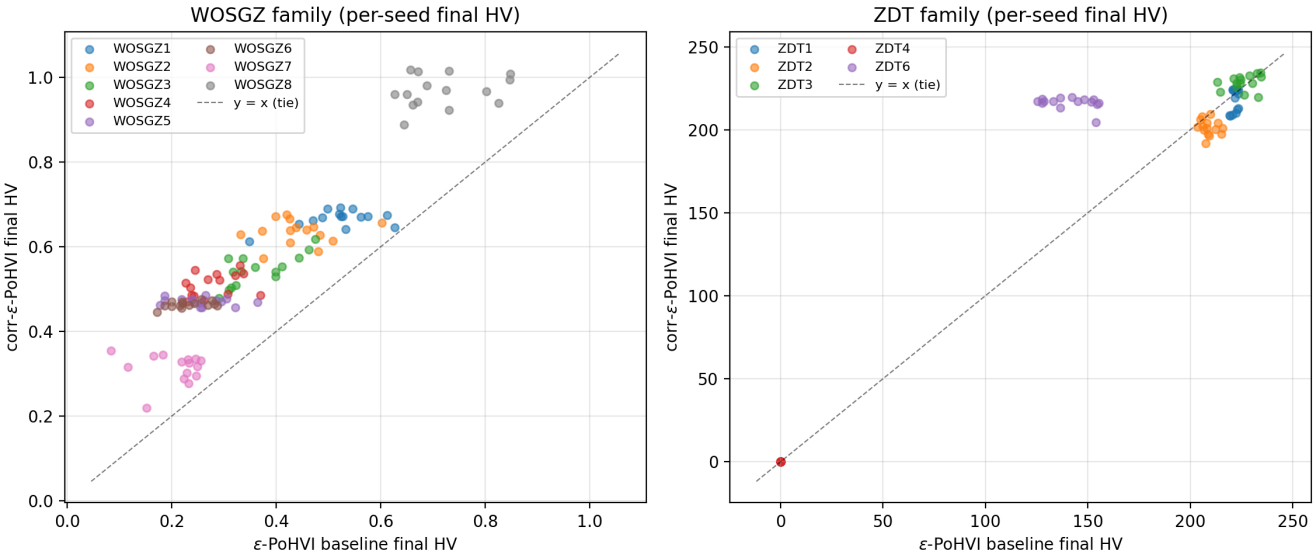

Figure 12: Per-seed final HV at $\boldsymbol{r}_{\text{acq}}$ after 170 BO iterations: ε-PoHVI baseline on the x-axis, corr-ε-PoHVI on the y-axis. Points above the $y = x$ diagonal are seeds where the correlated method wins. *Left:* WOSGZ family. *Right:* ZDT family.

Limitations. Several limitations of this comparison are worth stating explicitly:

- Only one MOGP architecture (ICM rank 1) is plugged into the correlated acquisition. The surrogate-selection chapter retains LMC and BoTorch MTGP as ablation alternatives; their full BO behavior is not characterized here.
- The benchmark is $m = 2$ objectives and $d = 30$ inputs. Generalization to higher-dimensional inputs or to $m \geq 3$ objectives is not tested.
- **Constrained BO benchmark** ($m = 2$) is out of scope: the constrained ε-CPoHVI distribution is derived in Chapter 4 and verified in Section 6.3, but is not run inside a full BO loop in this thesis. The natural future-work targets and rationale are in Section 6.3.3.
- **Generalization to $m \geq 3$ objectives** is a separate scope limitation (applying to the unconstrained method as well): the bivariate-Gaussian rectangle probability inside the cell-partition framework does not extend cleanly to higher-dimensional rectangle probabilities.
- The WOSGZ comparison cannot be made directly against Wang et al. [27] Table 4 due to the f_1-scaling divergence in their released ICML-rebuttal code; the paired-by-seed comparison within a canonical implementation used here sidesteps that issue but means absolute-magnitude comparisons against the published table should not be drawn.
- The acquisition optimizer is CMA-ES with BIPOP and 3 restarts on a 30-dimensional surface. Whether the correlated acquisition's landscape is more multimodal than the baseline's, or whether the 3-restart budget misses the global maximum at differential rates between the two methods, is not separately characterized here; an acquisition-landscape diagnostic is left to future work.

- The ε_t schedule ($\varepsilon_0 = 0.05$, $c = 0.02$) is the schedule of Wang et al. [27] and is not tuned in this thesis; sensitivity to this schedule is therefore a possible confound for both methods, but since both methods see the identical schedule it does not differentially affect the comparison.
- The BO trajectory $|\bar{\rho}|$ values reported in Figure 7 are averaged over the candidate points the optimizer actually evaluated; they are not a direct measurement of the $|\rho|$-weighted contribution to the acquisition surface itself, only of $|\rho|$ at the surface minima the optimizer found.

9 Conclusion

This thesis set out to extend the distribution-based hypervolume-improvement framework of Wang et al. [27] along the two axes the original authors identified as open problems: correlated objective posteriors and inequality-constrained search. Each extension is supported by a derivation, a numerical verification, and where the constraint scope allows, a full empirical comparison.

9.1 Summary of Contributions

Contribution 1: HVI distribution under correlated Gaussian objectives. Chapter 3 derives the exact probability density and cumulative distribution function of the generalized hypervolume improvement when the bivariate posterior carries arbitrary correlation $\rho \in (-1, 1)$. The cell-probability factor of Wang et al. is replaced by a bivariate-normal rectangle probability; the within-cell product density retains a Laurent-polynomial structure in the inner integration variable, which preserves the one-dimensional quadrature recipe used in the independent case. A consistency theorem verifies that every formula reduces exactly to the Wang et al. result at $\rho = 0$, so the correlated framework is a strict extension rather than a parallel construction.

Contribution 2: Constrained HVI distribution. Chapter 4 defines the constrained generalized HVI random variable $\Delta_C = \Delta(\boldsymbol{y}) \cdot \prod_j \mathbf{1}[g_j \leq 0]$ and derives its distribution in four progressively general settings: a single independent constraint, a single correlated constraint, multiple independent constraints, and multiple correlated constraints. The distribution is in every case a mixture of an infeasible point mass and a feasible continuous component, with the continuous component requiring the machinery of Contribution 1 because conditioning on the constraint values induces dependence between the objectives even when the unconditional posterior is independent. This yields a constrained ε-PoHVI (ε-CPoHVI) acquisition function that computes the probability of achieving at least ε *feasible* improvement, replacing in principle the standard product weighting of Gardner et al. [14]. The constrained acquisition is derived and numerically verified in Section 6.3 but is not benchmarked end-to-end against established constrained-MOBO baselines; that benchmark is left to future work (Section 9.4).

9.2 Empirical Findings

Numerical correctness. Both extensions are verified against 10^5-sample Monte Carlo simulation across 526 configurations summarized in Section 6.3, including 200 ICM-fitted posteriors drawn from ZDT1, ZDT2, and WOSGZ1-WOSGZ8. Empirical errors agree with Monte Carlo to within the expected sampling noise: the maximum absolute CDF error is 0.0059 (Case 2, unconstrained correlated), well below the 0.01 acceptance threshold; only 6 configurations exceed the per-configuration DKW 95% band of 0.0043 (versus ≈ 26 expected by chance under a perfect implementation), and zero configurations exceed the Bonferroni-adjusted family-wise band.

Bayesian-optimization performance. Across the 13 standard test problems, the correlated pipeline wins on every problem with exploitable inter-objective coupling, all eight WOSGZ problems and ZDT6, and is statistically indistinguishable from the baseline on three of the four remaining ZDT problems (ZDT1, ZDT3, ZDT4; the ZDT4 row is robust to the choice between the two

reference points used in Chapter 8 but is degenerate at the tight $\boldsymbol{r}_{\text{acq}}$, see Section 8.3), losing on one (ZDT2 at $p_{\text{Holm}} = 0.0085$). By family: **WOSGZ** 8/0/0 **(wins / ties / losses out of** 8 **problems)**, **ZDT** 1/3/1 **(out of** 5**)**, with no large-magnitude losses (ZDT2's -3.6% relative gap is the largest negative effect anywhere in the suite). On the WOSGZ family the relative HV gain ranges from $+28\%$ on WOSGZ1 to $+100\%$ on WOSGZ6; per-seed variance is 2-8$\times$ smaller for the correlated method, so the gain is in both per-iteration progress and run-to-run consistency.

Surrogate vs. acquisition. The empirical comparison varies the surrogate (independent GPs $\rightarrow$ ICM rank 1) and the acquisition (independent HVI distribution $\rightarrow$ correlated HVI distribution) jointly. The ICM-marginal ablation that would isolate the contribution of the correlated HVI distribution alone, using the ICM means and variances but discarding $\rho(\boldsymbol{x})$ in the acquisition, was not run in this thesis. The headline result therefore establishes that the combined ICM-plus-correlated-HVI pipeline outperforms the independent-GP-plus-independent-HVI baseline, not that the correlated HVI distribution alone is responsible for the gain. Two pieces of indirect evidence are consistent with the latter inference but do not establish it: the BO-trajectory $|\bar{\rho}|_{\text{BO}}$ values reported in Table 15 are non-trivial (0.12 to 0.23) on every problem the correlated pipeline wins and near-zero (10^{-4} to 3×10^{-3}) on the three ZDT problems where the comparison is otherwise clean (ZDT1, ZDT2, ZDT3); ZDT4 is the exception, with $|\bar{\rho}|_{\text{BO}} = 0.014$ but no win, discussed separately below. Second, the wins are largest in proportional terms where the cross-objective coupling is strongest. The cleanest follow-up is a single marginal-ablation experiment listed in Section 9.4.

Data efficiency, not just better asymptote. On the WOSGZ family the gain is best read as data efficiency rather than asymptotic improvement: the correlated pipeline reaches 90% of the best mean HV either method achieves in 65-113 BO iterations, while the baseline does not reach that target on any WOSGZ problem within the 170-iteration budget. On three problems (WOSGZ1, WOSGZ7, WOSGZ8) the paired gap modestly narrows in the late tail as the baseline catches up on mean accuracy, consistent with the surrogate-selection observation that ICM's RMSE advantage over independent GPs vanishes at $n = 230$. Whether the WOSGZ gap closes entirely at substantially larger n is not tested in this thesis.

When the method helps and when it does not. The win pattern broadly follows the BO-trajectory cross-objective signal $|\bar{\rho}|_{\text{BO}}$: WOSGZ problems sit at 0.12-0.23 and produce uniform 15/15-seed wins; ZDT1, ZDT2, ZDT3 sit at 10^{-4} to 3×10^{-3} and produce ties or a small loss; ZDT6 sits at the intermediate value 0.018 and produces a decisive $+53\%$ win. ZDT4 has comparable $|\bar{\rho}|_{\text{BO}}$ to ZDT6 (0.014) but does not win, indicating that the cross-objective signal is a necessary but not sufficient condition for an HV win: on ZDT4 the multimodal landscape and per-seed HV variance dominate the comparison even when the surrogate has some cross-objective signal to offer. The mechanism behind ZDT2 specifically losing where ZDT1 ties (both share the same lack of exploitable coupling) is not confidently attributed; candidate explanations (concave-front sensitivity, ICM's f_2 RMSE penalty at $n = 230$, sign-wrong fitted ρ at evaluated points) are discussed in Section 8.3.

9.3 Limitations

- **Surrogate vs. acquisition confound.** The BO comparison varies surrogate (independent GPs → ICM rank 1) and acquisition (independent HVI → correlated HVI) jointly. The ICM-marginal ablation, ICM means and variances combined with the independent HVI distribution by setting $\rho \equiv 0$ in the acquisition, was not run, so the experiment cannot apportion the WOSGZ gain between the better-calibrated MOGP surrogate and the correlated HVI distribution itself. This is the most directly addressable limitation and is listed first in Section 9.4.

- **Asymptote vs. data efficiency.** The BO comparison evaluates a 170-iteration budget. The paired WOSGZ HV gap visibly narrows in the late tail on at least three problems (WOSGZ1, WOSGZ7, WOSGZ8), suggesting the correlated pipeline buys faster convergence rather than a strictly better asymptote. Whether the gap closes entirely at substantially larger n is not tested.

- **Constrained acquisition derived but not benchmarked.** The constrained extension is verified against Monte Carlo simulation to within the expected sampling error (Section 6.3) but is not run inside a full BO loop in this thesis. The natural baselines (CEHVI, NEHVI-C / qNEHVI-C, MESMOC, PF^{2}ES) and test suites (C1/C2/C3-DTLZ, MW1-MW14) for such a comparison are well-established but assembling a fair benchmark across them is itself a research project; see Section 9.4.

- **Bi-objective only.** The cell-partition framework generalizes to $m \geq 3$ objectives in principle, but the bivariate Gaussian CDF used inside the correlated cell probability does not. Higher-dimensional rectangle probabilities require multivariate Gaussian integration, which is a non-trivial extension and applies to the unconstrained method as well as the constrained one.

- **Acquisition-optimization landscape diagnostic not run.** The acquisition optimizer is CMA-ES with BIPOP and 3 restarts on a 30-dimensional surface. Whether the correlated acquisition's landscape is more multimodal than the baseline's, or whether the 3-restart budget misses the global maximum at differential rates between the two methods, is not separately characterized here. Either would be a confound for the headline BO numbers.

- **Single MOGP architecture in the BO comparison.** Only GPyTorch ICM rank-1 was plugged into corr-ε-PoHVI for the headline comparison. Section 7 retains GPyTorch LMC rank-2 and BoTorch MultiTaskGP as alternative surrogates; their full BO behavior is not characterized here.

- **Synthetic benchmarks only.** Both ZDT and WOSGZ are synthetic test families. No real-world expensive function (materials design, hyperparameter tuning, simulation-based engineering) was used to validate the method outside the synthetic regime.

- **Fixed input dimensionality.** All BO experiments use $n_{\text{var}} = 30$. Whether the correlated method's advantage on WOSGZ-like problems persists at substantially higher input dimensionality is not tested.

- **No direct comparison against Wang et al. Table 4.** The WOSGZ implementation used here follows Wang, Ong, Sun, Gupta & Zhang [28] as published, whereas the released code of Wang et al. [27] uses an f_1-scaling variant that differs by a factor of 5 to 13× per point. Absolute-magnitude comparisons against the published baseline table should not be drawn; the comparison reported here is paired-by-seed within a canonical implementation.

9.4 Future Work

ICM-marginal ablation. The single experiment most directly motivated by the limitations of the present comparison is to plug the selected ICM rank-1 surrogate into the *independent* HVI distribution, using only the marginals of $\boldsymbol{\Sigma}(\boldsymbol{x})$ and discarding $\rho(\boldsymbol{x})$ in the acquisition, and rerun the 13-problem BO benchmark on the same paired-by-seed methodology used in Section 8. This three-way comparison (independent-GP / independent-HVI; ICM / independent-HVI; ICM / correlated-HVI) separates the contribution of the better-calibrated MOGP surrogate from the contribution of the correlated HVI distribution itself, and would settle the surrogate-vs-acquisition confound flagged in Section 9.3. The experiment is cheap relative to the original 13-problem run (same models, same trajectories, only the acquisition is changed) and is the highest-value follow-up.

Constrained-BO benchmark. A second direct continuation is to plug ε-CPoHVI into a full BO loop and benchmark it against the established constrained multi-objective acquisition baselines, CEHVI, NEHVI-C / qNEHVI-C in BoTorch [8], MESMOC, PF^{2}ES, on the standard constrained MOO test suites C1/C2/C3-DTLZ and MW1-MW14, with the same paired-by-seed methodology used in Section 8.

Higher-objective scaling. Extending to $m \geq 3$ objectives requires replacing the bivariate normal rectangle probability with an m-variate one. Approaches based on Genz quadrature or Monte Carlo with quasi-random sequences are the natural candidates; how their cost scales with m, and whether the cell-partition decomposition remains tractable, is an open question.

Decoupled and batch evaluation. The current framework assumes that all objectives and constraints are evaluated together at each candidate point. Decoupled evaluation, where each objective or constraint can be queried independently at different costs, is common in practice and would extend the framework into the domain of asynchronous and resource-aware BO. Similarly, deriving a batch q-ε-PoHVI [8] would extend the acquisition to parallel evaluation.

Adaptive ε schedules informed by the full distribution. The ε-schedule used here is the exponential decay of Wang et al. and is identical for both compared methods. With the full HVI distribution available, one could in principle pick ε adaptively per iteration based on quantiles of the predicted distribution, allowing the threshold to track the available improvement signal.

Real-world application. Finally, applying the method to a genuine expensive multi-objective problem with naturally correlated objectives (materials microstructure design, multi-fidelity simulation tuning, hyperparameter optimization under latency-accuracy trade-offs) would test whether the synthetic-benchmark advantage transfers to the regimes the framework was built for.

9.5 Closing Remarks

The two extensions developed in this thesis are independent in motivation but technically related: the constrained derivation requires the correlated machinery as a sub-routine, because conditioning on a shared constraint induces correlation in the objectives even when none was present in the unconditional posterior. The Wang et al. cell-partition framework therefore supports further extension, and distribution-based acquisition functions remain an active research direction. Across the 13 problems, the correlated pipeline produces decisive gains on every problem with substantial posterior cross-objective signal. On the two ZDT problems where the signal is at the numerical noise floor and the comparison is otherwise clean (ZDT1 and ZDT3), it ties the baseline. On the multimodal ZDT4 the comparison is dominated by other factors. On ZDT2 it produces a single small Holm-significant loss (-3.6%) whose mechanism is not confidently identified. This pattern is consistent with the broader research agenda of moving multi-objective Bayesian optimization away from moment-based acquisitions and toward distribution-based ones.

References

[1] Mauricio Álvarez and Neil D. Lawrence. Sparse convolved gaussian processes for multi-output regression. In *Advances in Neural Information Processing Systems*, volume 21, 2009.

[2] Mauricio A. Álvarez, Lorenzo Rosasco, and Neil D. Lawrence. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

[3] Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. BoTorch: A framework for efficient Monte-Carlo Bayesian optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[4] Syrine Belakaria, Aryan Deshwal, and Janardhan Rao Doppa. Max-value entropy search for multi-objective Bayesian optimization with constraints. In *NeurIPS 2020 Workshop on Machine Learning and the Physical Sciences*, 2020. arXiv:2009.01721.

[5] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[6] Edwin V. Bonilla, Kian Ming A. Chai, and Christopher K. I. Williams. Multi-task gaussian process prediction. In *Advances in Neural Information Processing Systems*, pages 153–160, 2007.

[7] Phillip Boyle and Marcus R. Frean. Dependent gaussian processes. In *Advances in Neural Information Processing Systems*, pages 217–224, 2004.

[8] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. In *Advances in Neural Information Processing Systems*, pages 2187–2200, 2021.

[9] Taco de Wolff, Alejandro Cuevas, and Felipe Tobar. MOGPTK: The multi-output Gaussian process toolkit. *Neurocomputing*, 424:49–53, 2021.

[10] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[11] Zvi Drezner and G. O. Wesolowsky. On the computation of the bivariate normal integral. *Journal of Statistical Computation and Simulation*, 35(1–2):101–107, 1990.

[12] Michael T. M. Emmerich, Kyriakos C. Giannakoglou, and Boris Naujoks. Single- and multiobjective evolutionary optimization assisted by gaussian random field metamodels. *IEEE Transactions on Evolutionary Computation*, 10(4):421–439, 2006.

[13] Paul Feliot, Julien Bect, and Emmanuel Vazquez. A Bayesian approach to constrained single- and multi-objective optimization. *Journal of Global Optimization*, 67(1):97–133, 2017.

[14] Jacob R. Gardner, Matt J. Kusner, Zhixiang Eddie Xu, Kilian Q. Weinberger, and John P. Cunningham. Bayesian optimization with inequality constraints. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pages 937–945. PMLR, 2014.

[15] Michael A. Gelbart, Jasper Snoek, and Ryan P. Adams. Bayesian optimization with unknown constraints. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 250–259. AUAI Press, 2014.

[16] Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2):141–149, 1992.

[17] Alan Genz and Frank Bretz. *Computation of Multivariate Normal and t Probabilities*, volume 195 of *Lecture Notes in Statistics*. Springer, 2009.

[18] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[19] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[20] Karl-Rudolf Koch. Multivariate normal distribution. In *Introduction to Bayesian Statistics*, chapter 2.5.1, pages 51–53. Springer, 2007.

[21] Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained Bayesian optimization with noisy experiments. volume 14, pages 495–519, 2019.

[22] Jonas Mockus. On bayesian methods for seeking the extremum. In *Optimization Techniques*, volume 27 of *Lecture Notes in Computer Science*, pages 400–404. Springer, 1974.

[23] Victor Picheny. A stepwise uncertainty reduction approach to constrained global optimization. *Journal of Global Optimization*, 64(1):25–44, 2016.

[24] Jixiang Qing, Henry B. Moss, Tom Dhaene, and Ivo Couckuyt. PF^{2}ES: Parallel feasible Pareto frontier entropy search for multi-objective Bayesian optimization under unknown constraints. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pages 2565–2588, 2023.

[25] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006.

[26] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016.

[27] Hao Wang, Kaifeng Yang, and Michael Affenzeller. Probability distribution of hypervolume improvement in bi-objective bayesian optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

[28] Zhenkun Wang, Yew-Soon Ong, Jianyong Sun, Abhishek Gupta, and Qingfu Zhang. A generator for multiobjective test problems with difficult-to-approximate Pareto front boundaries. *IEEE Transactions on Evolutionary Computation*, 23(4):556–571, 2019.

[29] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[30] Kaifeng Yang, Michael Emmerich, André Deutz, and Thomas Bäck. Efficient computation of expected hypervolume improvement using box decomposition algorithms. *Journal of Global Optimization*, 75(1):3–34, 2019.

[31] Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation*, 8(2):173–195, 2000.

[32] Eckart Zitzler and Lothar Thiele. Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999.

A Pilot Studies for the MOGP Benchmark

The surrogate-model benchmark (Section 7) fixes per-model Adam iteration budgets and restart counts based on two pilot studies. Both pilots were run before the headline benchmark and are reproduced here so the design choices in the main text are auditable.

A.1 Adam Convergence

Question. At what iteration count does each Adam-trained MOGP flatten on the WOSGZ negative-log-marginal-likelihood surface? Adam is not guaranteed to converge within a fixed budget on a 30-dimensional likelihood, and a fair benchmark requires every model to reach an Adam iteration count past which extra iterations no longer matter.

Protocol. Each Adam-trained MOGP candidate (GPyTorch ICM rank 1, GPyTorch LMC rank 2, MOGPTK SM-LMC) was fit on WOSGZ1 at $n = 230$ across three seeds. Training loss was logged at iteration checkpoints $\{100, 200, 500, 1000, 2000\}$ and the median trajectory was taken across seeds. The iteration budget was set to the smallest checkpoint past which the median loss improved by less than 1%.

Result. ICM rank 1 plateaued by iteration 1000; LMC rank 2 and SM-LMC continued to descend through iteration 2000 and use that as their ceiling. BoTorch `MultiTaskGP` uses L-BFGS-B rather than Adam and is not part of this pilot; it was given the same wall-clock budget as the longest Adam-trained model. Raw checkpoint data is in `experiments/result/pilots/adam_convergence_pilot.json`.

A.2 Restart-Count Asymmetry and Sensitivity

Asymmetric restart counts. Each MOGP was fit with three random restarts of the full hyperparameter optimizer; the independent-GP baseline used $5d = 150$ L-BFGS-B restarts, matching the configuration of Wang et al. [27]. These restart counts are asymmetric in favor of the baseline (the baseline is given 50× more restarts than every MOGP) making any MOGP win in the headline benchmark a conservative result.

Why ICM is robust to the small restart budget. A restart-count sensitivity check at ICM rank 1, run on WOSGZ1 at $n = 230$ across 10 seeds, compared $N \in \{3, 10\}$ restarts on identical training and test sets. $N = 10$ produced numerically identical predictions to $N = 3$ across all seeds: RMSE differences below 10^{-3} and NLPD differences below 10^{-2}. Per-restart initializations are independently drawn from the PyTorch random number generator, so the identical outcomes reflect a well-conditioned likelihood surface rather than a seeding artifact. Raw outputs are in `experiments/result/pilots/icm_restart_check.json`.

Why the baseline keeps the larger restart budget. The baseline's $5d = 150$ restart count is fixed by the reference configuration in Wang et al. [27]; reducing it to match the MOGPs would change the baseline rather than the comparison. We retain the original setting so that the ε-PoHVI

baseline in this thesis is the same algorithm, with the same hyperparameters, as the ε-PoHVI baseline of the published work.