Bayesian Prior Selection under Constraint-Based Feasible Regions

Parameter Learning in Bayesian Belief Networks under Data Scarcity

K.H.G. Keulen

Master of Science Thesis

TU Delft Delft University of Technology

Delft Center for Systems and Control

Bayesian Prior Selection under Constraint-Based Feasible Regions

Parameter Learning in Bayesian Belief Networks under Data Scarcity

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

K.H.G. Keulen

June 29, 2026

Faculty of Mechanical Engineering (ME) · Delft University of Technology



The work in this thesis was supported by TNO. Their cooperation is hereby gratefully acknowledged.





Copyright © Delft Center for Systems and Control (DCSC)
All rights reserved.

Abstract

Bayesian Belief Networks (BBNs) are widely used for reasoning under uncertainty in domains such as intelligence analysis. A representative example is the Hypothesis Management Framework (HMF), a BBN design pattern that assesses competing hypotheses from observed evidence, where decisions must be made from limited and incomplete data. In such settings, the accuracy of probabilistic inference depends strongly on the estimation of Conditional Probability Table (CPT) parameters. However, under severe data scarcity, conventional parameter-learning approaches either produce unstable estimates from limited observations, require a suitable source domain for transfer, or leave residual parameter uncertainty unrepresented.

This thesis develops and evaluates a feasible-region-based framework for prior construction and parameter learning in BBNs under limited data. Target-side uncertainty is represented as a feasible parameter region obtained by combining Wilson confidence intervals with qualitative domain-informed constraints. This feasible region forms the basis for two compatibility metrics, feasible-region probability mass and expected constraint violation, which quantify the alignment between candidate priors and the admissible parameter space. In an evaluation mode, the metrics assess transferred source knowledge; in a construction mode, they are combined into a cost function that builds informative priors directly from target-side information without requiring source-domain data. In addition, they provide interpretable diagnostics that identify the specific CPT rows and structural relationships responsible for prior–target mismatch.

The framework is evaluated through a controlled simulation study on a minimal HMF network. Results show that the proposed compatibility metrics successfully distinguish between source priors of varying quality and provide localized diagnostics of incompatibility. Building on these metrics, the construction-mode cost function combines feasible-region mass and constraint violation with a concentration regularizer, constructing Beta priors that balance feasible-region coverage, structural consistency, and calibrated uncertainty. Across all tested target sample sizes and constraint settings, the optimized priors reduce CPT posterior mean absolute error by 14–56% relative to a uniform prior. The method consistently outperforms moderate and conflicting transfer-learning baselines and remains competitive with a well-aligned source-domain prior despite requiring no external data. When propagated through network inference, these parameter-level improvements amplify, yielding network-query error reductions of up to 66%.

The results demonstrate that target-side uncertainty and qualitative domain knowledge can be transformed into informative Bayesian priors that improve parameter learning and downstream inference under severe data scarcity. The proposed method therefore provides a practical alternative in settings where suitable source domains are unavailable, difficult to validate, or susceptible to negative transfer.

Table of Contents

Acknowledgements **xv**

1 Introduction **1**

1-1 Motivation . . . 1

1-2 Background information . . . 2

1-2-1 Bayesian Belief Networks in decision support . . . 2

1-2-2 Risk of transferring incorrect prior knowledge . . . 3

1-3 Problem description and gaps . . . 4

1-4 Thesis objective and research questions . . . 5

1-5 Contribution and thesis outline . . . 6

2 Background and related work **9**

2-1 Bayesian Belief Networks . . . 9

2-1-1 Definition and parameterization . . . 9

2-1-2 Maximum likelihood estimation . . . 10

2-2 Parameter learning under data scarcity . . . 11

2-2-1 Limitations of ML estimation . . . 11

2-2-2 Bayesian estimation and prior specification . . . 11

2-2-3 Limitations of prior-based estimation . . . 12

2-2-4 Interval estimation under data scarcity . . . 13

2-3 Qualitative constraints and constrained parameter learning . . . 14

2-3-1 The role of qualitative constraints . . . 14

2-3-2 Constrained learning methods . . . 15

2-3-3 Parameter spaces under constraints . . . 15

2-4 Transfer learning for BBN parameter learning: MPL-TC . . . 16

2-5 CPT uncertainty propagation in BBNs . . . 17

2-6 The Hypothesis Management Framework . . . 17

2-7 Summary . . . 20

3 Methodology **23**
3-1 Problem setting and assumptions . . . 24
3-1-1 Prior design setting . . . 24
3-1-2 Modeling assumptions . . . 24
3-2 Feasible parameter region from target information . . . 25
3-2-1 Representation of target uncertainty . . . 25
3-2-2 Qualitative cross-row constraints . . . 27
3-2-3 Joint feasible-region construction . . . 27
3-3 Metrics for prior-region compatibility . . . 28
3-3-1 Prior mass in the feasible region . . . 29
3-3-2 Constraint violation metrics . . . 29
3-4 Source domain prior evaluation . . . 31
3-4-1 Source prior representation . . . 31
3-4-2 Compatibility-based source prior ranking . . . 31
3-4-3 Limitations of compatibility-based evaluation . . . 32
3-5 Diagnostic violation decomposition . . . 32
3-5-1 Decomposition of global incompatibility . . . 32
3-5-2 Identifying critical CPT parameters . . . 33
3-6 Bayesian prior optimization . . . 33
3-6-1 Parametric prior family . . . 34
3-6-2 Distance to representative target configuration . . . 35
3-6-3 Base cost function . . . 36
3-6-4 Target-uncertainty regularization of prior concentration . . . 37
3-6-5 Combined cost function . . . 38
3-7 Optimization procedure . . . 38
3-7-1 Monte Carlo approximation of the cost function . . . 39
3-7-2 Derivative-free optimization . . . 39
3-7-3 Multi-start strategy . . . 40
3-8 Inference with optimized priors . . . 40
3-9 Validation design . . . 41
3-9-1 Reproducibility . . . 41
3-9-2 Evaluation metrics . . . 41
3-9-3 Query posterior uncertainty characterization . . . 43
3-10 Summary . . . 43

4 Experimental setup and case study **45**
4-1 Experimental setup . . . 45
4-1-1 Simulated BBN and target model . . . 45
4-1-2 Feasible-region construction . . . 47
4-1-3 Experimental modes . . . 48
4-1-4 Evaluation mode: Source prior families for compatibility metric validation . . . 48
4-1-5 Optimization mode: Methods compared for prior optimization and inference . . . 49
4-1-6 Experimental design and performance metrics . . . 50

4-2 Compatibility metric validation . . . 51
4-2-1 Evaluation procedure . . . 51
4-2-2 Row-wise vs joint compatibility . . . 52
4-2-3 Ranking and sensitivity to constraint tightness . . . 52
4-2-4 Robustness across seeds . . . 54
4-2-5 Incompatibility localization . . . 54
4-3 Prior optimization and Parameter Learning . . . 55
4-3-1 Optimization configuration . . . 56
4-3-2 CPT posterior accuracy . . . 57
4-3-3 Comparison with MPL-TC . . . 60
4-4 Network-level Inference Evaluation . . . 64
4-4-1 Experimental setup . . . 64
4-4-2 Query posterior accuracy . . . 65
4-4-3 Per-query analysis . . . 66
4-5 Discussion . . . 69
4-6 Summary . . . 70

5 Discussion and Conclusion 73
5-1 Overall Thesis Conclusion . . . 73
5-2 RQ1: Feasible-region construction . . . 74
5-2-1 Result . . . 74
5-2-2 Insights . . . 74
5-2-3 Limitations . . . 74
5-3 RQ2: Compatibility quantification and decomposition . . . 75
5-3-1 Result . . . 75
5-3-2 Insights . . . 75
5-3-3 Limitations . . . 76
5-4 RQ3: Prior construction and downstream inference . . . 76
5-4-1 Result . . . 76
5-4-2 Insights . . . 77
5-4-3 Limitations . . . 77
5-5 Conclusion and future work . . . 78
5-5-1 Directions for future work . . . 79

A Simulation-based experiment 81
A-1 Example derivation of the query $P(H_1 = \mathrm{T} \mid I = \mathrm{Present})$ . . . 81

B Supplementary Results for Simulation Experiments 83
B-1 Feasible region construction and compatibility metrics . . . 83
B-1-1 Feasible region projections . . . 83
B-1-2 Compatibility metric validation . . . 83
B-1-3 Incompatibility localization . . . 86
B-2 Prior optimization . . . 88
B-2-1 Computational settings . . . 88
B-2-2 Optimization behavior . . . 89

B-3 CPT parameter-learning results . . . 89
B-3-1 Representative feasible-region-based optimization result . . . 89
B-3-2 Full CPT comparison tables . . . 90
B-3-3 Statistical Significance Tests . . . 91
B-4 Network-level inference results . . . 92
B-4-1 Complete Network performance . . . 92
B-4-2 Query-level uncertainty analysis . . . 93

Glossary **105**
List of Acronyms . . . 105
List of Symbols . . . 105

List of Figures

1-1 Example BBN for the burglary-alarm scenario, illustrating the DAG structure and conditional dependency relations among the variables. 2

1-2 Bayesian transfer learning constructs an informative prior from source-domain knowledge, pulling the posterior toward the source estimate. When the source and target domains are well-matched this improves estimation, but when they differ the posterior is pulled away from the true target parameters. Adapted from [1]. 4

2-1 The burglary-alarm BBN. Each node is associated with a local CPT. The row $(B{=}\mathsf{T}, E{=}\mathsf{F})$ specifies $\mathbb{P}(A{=}\mathsf{T} \mid B{=}\mathsf{T}, E{=}\mathsf{F}) = 0.94$. JohnCalls and MaryCalls are conditionally independent of Burglary and Earthquake given the Alarm state. 10

2-2 Effect of prior concentration on posterior updating given the same observed data. A high-concentration prior resists updating and keeps the posterior near the prior mean; a low-concentration prior shifts readily toward the likelihood [2]. 13

2-3 Graphical model representation of MPL-C with M constraints [3]. 16

2-4 A non-disjoint HMF model. Hypotheses, indicators, and statements form three distinct layers with directed edges encoding diagnostic relationships. 18

2-5 Schematic representation of a disjoint model (left) and a non-disjoint model (right) [4]. 19

2-6 Small HMF network with two hypotheses, one indicator, and one statement. 20

3-1 Geometric illustration of the two violation components in Equation $(3\text{-}12)$ for a two-row CPT. The credal box $\mathcal{B}_i$ (blue) is defined by the row-wise Wilson intervals. The feasible region $\mathcal{F}_i$ (green) is the intersection of $\mathcal{B}_i$ with the monotonicity constraint $\theta_{FT} \geq \theta_{FF}$. A parameter vector θ outside $\mathcal{F}_i$ incurs box violations $V_{\mathrm{box},j}$ measuring how far θ extends beyond each Wilson interval bound (orange), and a cross-row constraint violation $V_{c,\ell} = \max(\theta_{FF} - \theta_{FT}, 0)$ equal to the algebraic constraint surplus (purple). 30

3-2 Overview of the proposed prior optimization framework. Domain constraints and limited target data are combined into the feasible region $\mathcal{F}_i$ (Section 3-2). Compatibility metrics are incorporated into a cost function minimized by Powell's method with multiple restarts (Sections 3-6–3-7), producing an optimized Beta prior for each row of the indicator CPT $\mathbb{P}(I \mid H_1, H_2)$. Dashed lines mark the true parameter values θ^*. The optimized priors are updated conjugately with target inference data to yield CPT posteriors, propagated via belief propagation to hypothesis query posteriors. The highlighted node I in the HMF network indicates where prior optimization is applied. 39

4-1 Simulated HMF network used throughout the simulation experiment. The primary object of analysis is the CPT $\mathbb{P}(I \mid H_1, H_2)$, consisting of four rows corresponding to the four binary parent configurations. 46

4-2 Prior mass in the feasible region as a function of target sample size for each source family, under no constraints (left), mild constraints (center), and strict constraints (right). Higher values indicate better compatibility with the target-side feasible region. The conflict source achieves near-zero mass under both constraint modes. Discriminative power between intermediate families increases substantially as constraints tighten. Shaded bands indicate ± 1 standard deviation across seeds. 53

4-3 Mean expected violation of the feasible region as a function of target sample size for each source family, under no constraints (left), mild constraints (center), and strict constraints (right). Lower values indicate better consistency with the target-side constraints. The conflict source exhibits consistently high violation across all settings. Shaded bands indicate ± 1 standard deviation across seeds. 54

4-4 Row-wise box violation heatmap for all source families at $N = 50$, strict constraints, averaged over 20 seeds. Each cell reports the mean violation of that CPT row against its Wilson interval bound. The conflict source concentrates violation almost entirely on θ_{FF} (0.531), indicating that its main row-wise mismatch is the inflated baseline probability assigned to the (F, F) parent configuration. The remaining rows show substantially smaller violations. The equal, very mild, and mild shift sources show near-zero row-wise violation, while the strong shift source produces a moderate but more evenly distributed violation profile. 55

4-5 Constraint-level violation rates for four source families at $N = 50$, strict constraints. The conflict source (red) exhibits the highest violation rates across most constraints, with complete violation of the gap constraint C_7 ($\theta_{TT} \geq \theta_{FF} + 0.30$), followed by substantial violations of the monotonicity constraints C_1 ($\theta_{FF} \leq \theta_{FT}$), C_4 ($\theta_{TF} \leq \theta_{TT}$), and C_2 ($\theta_{FF} \leq \theta_{TF}$). These constraints are directly affected by the inflated θ_{FF} value identified in the row-wise heatmap. The equal source shows a notable violation on the closeness constraint C_5 ($\theta_{FT} \leq \theta_{TF} + 0.15$), indicating that θ_{FT} and θ_{TF} lie near the boundary of the allowed tolerance band even when individual rows remain consistent with their Wilson intervals. 56

4-7 Posterior means and 95% credible intervals per CPT row for the representative seed ($N = 50$, strict constraints, seed 12). Ground-truth parameters are shown as dashed lines. The optimized prior reduces both posterior error and credible interval width across all four rows. 57

4-6 Prior and posterior Beta distributions per CPT row for the representative seed ($N = 50$, strict constraints, seed 12). The shaded region indicates the target-side Wilson interval. The optimized prior concentrates probability mass within the constraint-consistent feasible region, shifting the posterior toward the ground-truth parameter (dashed vertical line) relative to the uniform prior. 58

4-8 Posterior MAE as a function of target sample size for all methods, faceted by constraint mode. Shaded bands indicate ± 1 standard deviation across 15 seeds. The optimized prior (red) consistently achieves lower MAE than the uniform baseline (orange) at all N. The MPL-TC conflict source (dashed purple) remains the worst-performing method across the full N range. 61

4-9 Bernoulli KL divergence from the true CPT parameters to the posterior mean estimate, as a function of target sample size, faceted by constraint mode. The MPL-TC mild source (green) exhibits higher KL than the uniform prior at small N despite achieving lower MAE, reflecting systematic source bias. The optimized prior (solid red) achieves the lowest KL throughout $N \leq 100$, indicating distributional accuracy. Shaded bands show the 10th–90th percentile range across 15 seeds rather than ± 1 standard deviation, since KL divergence is non-negative and right-skewed. 62

4-10 MPL-TC source-size sensitivity. Each panel shows CPT posterior MAE vs target sample size for one source quality level (mild, moderate, conflict), with separate lines for each source dataset size N_{src}. The optimized prior (solid red) and uniform baseline (dashed orange) are shown as references; neither uses any source data. Shaded bands indicate ± 1 standard deviation across 15 seeds. The optimized prior uses no source data at any point. 63

4-11 Average network query MAE as a function of target sample size, under both constraint modes. The optimized prior is shown for both constraint modes. The uniform baseline (dashed orange) and three MPL-TC sources (mild, moderate, conflict) are also shown. Shaded bands indicate ± 1 standard deviation across 15 seeds. 65

4-12 Bernoulli KL divergence from the true network query posteriors to the estimated posteriors, as a function of target sample size. MPL-TC sources are aggregated over constraint modes; the optimized prior is shown separately for mild (solid red) and strict (dash-dot dark red) constraints. Error bars indicate ± 1 standard deviation across 15 seeds. 66

4-13 Mean absolute query error per inference query at $N = 50$, strict constraints, averaged over 15 seeds. Error bars indicate ± 1 standard deviation across seeds. 67

4-14 Query posterior distributions represented as KDE curves, pooled over 15 seeds at $N = 50$, strict constraints. Each panel shows one of the five inference queries. The dashed vertical line indicates the true query posterior under the ground-truth BBN. The optimized posterior (red) is more concentrated and its mean is consistently closer to the true value than the uniform posterior (orange), reflecting reduced parameter uncertainty and improved distributional placement. The full comparison including all MPL-TC sources is shown in Appendix B-4-2. 71

B-1 2D projection of the feasible region $\mathcal{F}$ onto the $(\theta_{FF}, \theta_{FT})$ plane for a representative experiment. The outer rectangle corresponds to the Wilson-score credal box $\mathcal{B}$, while the shaded region denotes the subset satisfying all cross-row constraints. The projection illustrates how qualitative monotonicity and closeness constraints reduce the admissible parameter space relative to the unconstrained uncertainty intervals. . 84

B-3 Source family ranking stability across 20 random seeds, under strict constraints at $N = 100$. Each cell shows the frequency with which a source family achieves a given rank position by prior mass in the feasible region, across seeds. Darker cells indicate more consistent ranking. The equal source is top-ranked and the conflict source is bottom-ranked in the large majority of seeds, while intermediate families show greater variability. 84

B-2 2D projection of the feasible region $\mathcal{F}$ onto the $(\theta_{FF}, \theta_{TT})$ plane for the same experiment. The feasible region forms a structured subset of the Wilson-score credal box due to the imposed cross-row constraints. This projection highlights the dependence introduced between parameters that would otherwise be represented independently through row-wise uncertainty intervals. 85

B-4 Mean compatibility base cost function (mass + violation + distance) value for the four representative source priors under strict constraints at $N = 50$, averaged over 20 seeds. Lower values indicate greater compatibility with the feasible region. The matched and mild-shift priors achieve the lowest cost values, while the conflict prior is consistently assigned the highest penalty. Error bars denote one standard deviation across seeds. The ranking demonstrates that the base cost function successfully discriminates between increasing levels of prior-target mismatch. 86

B-5 Mean violation decomposition across six source prior families under strict constraints ($N = 50$). Each bar separates the total mean violation into box violation V_{box} (prior mean outside the Wilson calibration interval) and structural violation V_c (cross-row constraint violations). The equal source produces near-zero violation on both components. Conflict produces dominant structural violation from monotonicity reversal, while strong shift accumulates primarily box violation from location mismatch. Mild and moderate sources show mixed contributions, illustrating that equal total violation can arise from qualitatively different structural sources. 87

B-6 Row-wise box violation heatmap for all source families at $N = 50$, mild constraints, averaged over 20 seeds. Each cell reports the mean violation of that CPT row against its Wilson interval bound. The conflict source concentrates violation almost entirely on θ_{FF} (0.513), indicating that its main row-wise mismatch is the inflated baseline probability assigned to the (F, F) parent configuration. The remaining rows show substantially smaller violations. The strong shift source produces a more moderate violation profile. 87

B-7 Representative optimization result for a single experimental run ($N = 50$, strict constraints, seed 12). The shaded region denotes the feasible parameter set $\mathcal{F}$ obtained from the Wilson-score intervals and cross-row monotonicity constraints. The projected center of the feasible region serves as the reference point in the distance component of the cost function, while the optimized prior concentrates probability mass within the admissible region. The figure illustrates how the optimization balances feasible region coverage, constraint satisfaction, and proximity to the representative interior of $\mathcal{F}$. 90

B-8 Paired Wilcoxon signed-rank test results under strict constraints, for all comparisons (optimized prior vs uniform, MPL-TC mild, moderate, and conflict) across all target sample sizes N and three metrics: CPT posterior MAE (left), CPT KL divergence (centre), and network query MAE (right). $n = 15$ seeds per group. Green cells indicate a statistically significant advantage of the optimized prior over the baseline ($p < 0.05$); red cells show the p-value where significance is not reached. The optimized prior significantly outperforms all baselines at $N \leq 30$ on CPT MAE and network MAE. The KL divergence advantage over MPL-TC mild is maintained through $N = 100$, after which both methods converge. The conflict source comparison remains significant across the full N range on all three metrics. 92

B-9 Paired Wilcoxon signed-rank test results under mild constraints, using the same layout as Figure B-8. $n = 15$ seeds per group. Results are consistent with strict constraints: the optimized prior significantly outperforms all baselines at $N \leq 30$ on CPT MAE and network MAE, the KL advantage over MPL-TC mild is maintained through $N = 100$, and the conflict source comparison remains significant throughout on all three metrics. 92

B-10 Full comparision of query posterior distributions represented as KDE curves, pooled over 15 seeds at $N = 50$, strict constraints. Each panel shows one of the five inference queries. The dashed vertical line indicates the true query posterior under the ground-truth BBN. The optimized posterior (red) is more concentrated and its mean is consistently closer to the true value than the uniform posterior (orange) or the MPL-TC source posteriors, reflecting reduced parameter uncertainty and improved distributional placement. 95

List of Tables

4-1 Ground-truth CPT parameters for $\mathbb{P}(I = \text{Present} \mid H_1, H_2)$ in the simulated target model. 46

4-2 Source prior families used for compatibility metric validation, ordered by increasing mismatch from the target. Values give $\mathbb{P}(I = \text{Present} \mid H_1, H_2)$ for each parent configuration. Target values: $(\theta_{FF}, \theta_{FT}, \theta_{TF}, \theta_{TT}) = (0.10,\ 0.70,\ 0.60,\ 0.90)$. . . . 49

4-3 MPL-TC source quality levels used in the parameter-learning benchmark. Values give $\mathbb{P}(I = \text{Present} \mid H_1, H_2)$ for each parent configuration. Target: $(0.10,\ 0.70,\ 0.60,\ 0.90)$. 50

4-4 Factorial design of the simulation experiment. Each combination of seed, N_{target}, and constraint mode constitutes one experimental run. 50

4-5 Expected ranking of source families by compatibility metric, from most compatible (rank 1) to least compatible (rank 6). 51

4-6 Prior mass in the feasible region and mean expected violation for each source family across all target sizes, under strict constraints ($\varepsilon = 0.15$), averaged over 20 seeds. The rows ordered by prior mass at $N = 75$, under which the strong-shift source ranks above the moderate source at small to moderate target sizes despite its larger nominal deviation, because the moderate source places more mass on constraint-violating configurations and is penalized more heavily by the feasible-region geometry. 52

4-7 Comparison of optimization configurations for prior construction. Best-performing configuration highlighted in bold. 56

4-8 Per-row inference results for the representative seed at $N = 50$, strict constraints (seed 12). $[L_j, U_j]$ is the Wilson calibration interval; $[a_j, b_j]$ is the marginal projection of $\mathcal{F}_{i,\text{mode}}$ after cross-row constraint intersection. τ denotes the concentration of the optimized Beta prior. 58

4-9 Summary of optimized prior performance relative to the uniform baseline across all target sizes and both constraint modes, averaged over 15 seeds. MAE improvement is the per-seed percentage $(\text{MAE}_{\text{uni}} - \text{MAE}_{\text{opt}})/\text{MAE}_{\text{uni}} \times 100\%$, averaged over seeds. Coverage is the proportion of CPT rows for which the true θ^* falls within the 95% posterior credible interval, reported for both the uniform (uni) and optimized (opt) priors. 59

4-10 CPT posterior MAE under strict constraints across all target sizes, averaged over 15 seeds. Improvement percentages are computed relative to the optimized prior; negative values indicate the comparison method outperforms the optimized prior. . . 60

4-11 Network query MAE under strict constraints across all target sizes, averaged over 15 seeds. Improvement percentages are computed relative to the optimized prior; negative values indicate the comparison method outperforms the optimized prior. . . 65

4-12 Mean absolute query error per inference query at $N = 50$, strict constraints, averaged over 15 seeds. 67

4-13 Mean decision error rate (proportion of queries with incorrect binary classification at threshold 0.5) under strict constraints, averaged over 15 seeds. Bold indicates the lowest error at each N. 68

B-1 Computational settings for the prior optimization experiments. 88

B-2 Mean optimized concentration parameters $\tau_j = \alpha_j + \beta_j$ per CPT row, averaged over 15 seeds. Values in parentheses are standard deviations. Strict constraints systematically inflate τ_{FT} and τ_{TF} relative to mild constraints, while τ_{TT} remains lowest due to sparse calibration data in the (T, T) conditioning event. 89

B-3 CPT posterior MAE for mild and strict constraints across all target sizes, averaged over 15 seeds. Baseline columns are identical across constraint modes. The optimized prior achieves lower MAE under mild constraints throughout, as the less restrictive feasible region allows greater flexibility in prior placement. 91

B-4 CPT posterior inference performance across all target sizes and both constraint modes, averaged over 15 seeds. MAE and RMSE improvements are per-seed percentage improvements averaged over 15 seeds. CI reduction is the mean reduction in 95% posterior credible interval width. KL is the mean Bernoulli KL divergence from the true parameter to the posterior mean. Coverage is the proportion of CPT rows for which θ^* falls within the 95% posterior CI. 91

B-5 Network query MAE for mild and strict constraints across all target sizes, averaged over 15 seeds. Baseline columns (Uniform and MPL-TC sources) are identical across constraint modes because those priors are constructed independently of the feasibility constraints. The optimized prior MAE differs between modes, as strict constraints restrict the feasible region further. 93

B-6 Network query MAE and mean Bernoulli KL divergence $\mathrm{KL}(p_{\mathsf{true}} \| p_{\mathsf{est}})$ for mild and strict constraints, averaged over 15 seeds and all five queries. Baseline columns are identical across constraint modes.The optimized prior achieves the lowest or near-lowest MAE and KL in most regimes, while consistently outperforming the uniform and conflict baselines; MPL-TC mild yields lower MAE than uniform at small N but higher KL, reflecting that a concentrated but misaligned prior inflates KL more than it inflates MAE. 93

Acknowledgements

First, I would like to thank my TU Delft supervisor, Dr. Ir. N. J. (Nitin) Myers, for his guidance and support throughout this thesis. I am grateful for the valuable insights, critical feedback, and openness to engaging discussions that helped shape both the direction and quality of this research. His enthusiasm for the subject and willingness to challenge ideas encouraged me to think more deeply about the problems addressed in this research.

I would also like to thank my TNO supervisor, Ir. L. (Laura) Middeldorp, for her continuous support throughout the project. I greatly appreciated her enthusiasm, practical insights, and willingness to make time for questions and discussions whenever needed. Her perspective helped bridge the gap between academic research and real-world application. Our regular discussions and her constructive feedback were invaluable, both for the development of this research and for my personal and professional growth.

In addition, I would like to thank everyone at TNO for warmly welcoming me and making me feel at home within the department. The openness, support, and pleasant working environment contributed greatly to a positive and memorable internship experience.

Besides this, I am also grateful to those closest to me for their unwavering support during this journey.

Finally, this thesis marks the end of my time at TU Delft. Looking back, I am grateful for the experiences, knowledge, friendships, and opportunities that the university has offered me over the past years. The journey has been both challenging and rewarding, and I look forward to applying the lessons learned during this period to the next stage of my career.

Delft University of Technology K.H.G. Keulen June 29, 2026

Chapter 1

Introduction

1-1 Motivation

In modern intelligence operations, the ability of decision makers to make accurate, prompt, and informed choices is critical to gaining strategic advantage. Effective decision-making hinges on achieving a superior information position and requires more than access to data. It depends on two complementary capabilities: situational awareness, the ability to perceive and interpret what is currently observed, and situational understanding, the ability to grasp the implications of those observations by uncovering underlying causes and anticipating future developments. Together, these capabilities form the foundation of effective command and control. Achieving them requires analysts to draw coherent conclusions from incomplete, noisy, and rapidly evolving information, combining fragmentary evidence from multiple sources into reliable probabilistic assessments. Failure to do so can lead to missed escalations, delayed responses, or misallocation of resources with potentially serious consequences.

To support this kind of structured reasoning under uncertainty, Bayesian Belief Networks (BBNs) have been widely adopted across decision support applications [5]. BBNs represent probabilistic relationships between variables through a Directed Acyclic Graph (DAG) and associated Conditional Probability Tables (CPTs), enabling analysts to combine heterogeneous evidence and update beliefs as new information becomes available. The reliability of a BBN depends critically on the accuracy of these CPT parameters, which determine how strongly evidence influences competing hypotheses. For example, in the burglary-alarm network shown in Figure 1-1, a CPT row specifies the probability that the alarm sounds given that a burglary occurred while no earthquake took place:

$$\mathbb{P}(A = \text{True} \mid B = \text{True}, E = \text{False}).$$

Errors in such parameters directly affect the posterior probability assigned to the underlying hypothesis, and a poorly calibrated network can produce misleading assessments even when the evidence itself is reliable.

Learning accurate CPT parameters, however, requires sufficient observational data for each configuration of parent variables. Most existing learning algorithms assume large and complete datasets [6]. In many operational settings, this assumption does not hold: observations are sparse, incomplete, or collected under non-stationary conditions. This is particularly evident in analytical decision support systems such as the Hypothesis Management Framework (HMF),

which supports the analysis of competing hypotheses given partially observed evidence [7]. For example, an analyst assessing whether an ongoing demonstration is likely to escalate into violence must reason from a small number of observations, such as crowd size, confrontations with security forces, or reports of property damage, none of which may have been seen often enough to estimate their conditional relationships reliably. Under such data scarcity, standard Bayesian learning produces unstable parameter estimates that can distort downstream inference.

A natural response is to supplement scarce target data with knowledge transferred from a related domain. Transfer learning approaches such as Multinomial Parameter Learning with Transferred priors and Constraints (MPL-TC) [8] construct informative priors from source data, improving parameter estimates when source and target domains are sufficiently similar. However, transfer introduces a fundamental risk: when the source domain differs from the target, the transferred prior systematically biases posterior estimates away from the truth, a phenomenon known as negative transfer [1]. In practice, the degree of mismatch between source and target is rarely known in advance. At the same time, analysts often possess qualitative domain knowledge that could constrain the parameter space, for example the belief that the probability of an alarm sounding should always be higher when a burglary is occurring than when it is not, regardless of other conditions.

These difficulties, namely sparse observations, unreliable source transfer, and qualitative knowledge that resists formal encoding, highlight a gap in existing BBN parameter-learning methods. Current approaches either assume sufficient data, accept source knowledge without systematic evaluation, or incorporate qualitative constraints only as hard penalties during estimation. A method that explicitly represents target-side uncertainty, evaluates the compatibility of candidate prior knowledge, and constructs informative priors from limited data and qualitative domain knowledge alone remains lacking.

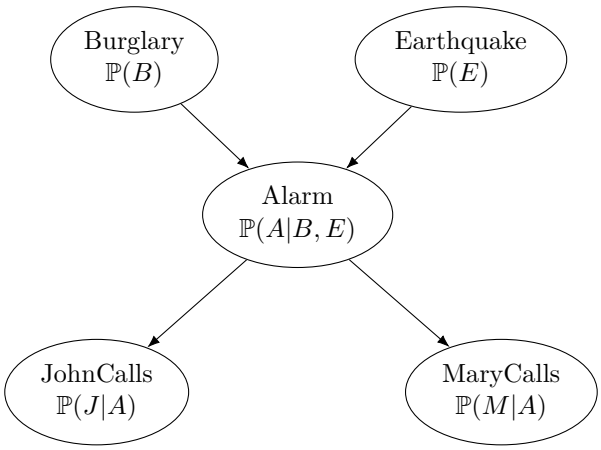


Figure 1-1: Example BBN for the burglary-alarm scenario, illustrating the DAG structure and conditional dependency relations among the variables.

1-2 Background information

1-2-1 Bayesian Belief Networks in decision support

A BBN is a probabilistic graphical model represented by a DAG, in which each node corresponds to a random variable and each directed edge encodes a conditional dependence. The burglary–alarm example introduced in Figure 1-1 illustrates such a network.

The model consists of:

- A qualitative component: the DAG structure, expressing (in)dependence relations. However, in this thesis the structure is assumed to be known and fixed.
- A quantitative component: a set of CPTs specifying the probability distribution of each node given its parents.

Random variables may be discrete or continuous, though this thesis focuses exclusively on discrete BBNs. For discrete variables, the CPT enumerates probabilities over finitely many states.

BBNs are particularly attractive in intelligence-support settings because they:

- Integrate heterogeneous information sources, including empirical data, expert judgment, and prior knowledge, making them suitable for domains characterized by uncertainty and limited observations [9, 10].
- Represent complex stochastic systems through an interpretable graphical structure that explicitly captures conditional dependencies between variables [11].
- Support probabilistic reasoning and evidence updating, and under appropriate structural assumptions, can additionally support causal interpretation and inference [12, 13].

In intelligence-support applications such as the HMF, BBNs are used to quantify the likelihood of competing hypotheses given incomplete and evolving evidence. However, when data is limited, incomplete, or noisy, constructing reliable CPTs becomes difficult. Analysts must often rely on expert judgment not only to quantify probabilities, but also to define structural dependencies between variables. This dual reliance on expert knowledge increases epistemic uncertainty and directly affects the reliability of downstream inference.

1-2-2 Risk of transferring incorrect prior knowledge

A natural response to data scarcity in BBN parameter learning is to transfer knowledge from related source domains. If a similar network has been learned in a different region, time period, or operational context, its parameters can serve as informative priors for the target network, as illustrated in Figure 1-2.

This idea underlies transfer learning approaches such as MatchCPT fitness with Aggregation fusion (CPTAgg) [14] and MPL-TC [8]. The approach is intuitively appealing because it leverages potentially informative source-domain knowledge when target observations are limited.

However, transfer introduces a fundamental risk. If the source and target domains are not sufficiently similar, the transferred prior can pull the posterior away from the true target parameters rather than toward them, as illustrated in Figure 1-2. This phenomenon is known as negative transfer [1]. In the BBN context, negative transfer manifests as systematic bias in posterior CPT estimates, which then propagates through the network and distorts downstream inference. Transfer learning approaches such as CPTAgg [14] combine target and auxiliary CPT estimates through similarity-weighted aggregation, improving performance when source and target are closely related. However, when source and target distributions differ substantially, such heuristic aggregation cannot guarantee that incompatible source knowledge is excluded from the update, and may still introduce negative transfer. The CPTAgg method, for example, has been shown to fail to outperform even a simple Maximum A Posteriori (MAP) estimate in some settings precisely because similarity-weighted aggregation breaks down when domain mismatch is non-trivial [8].

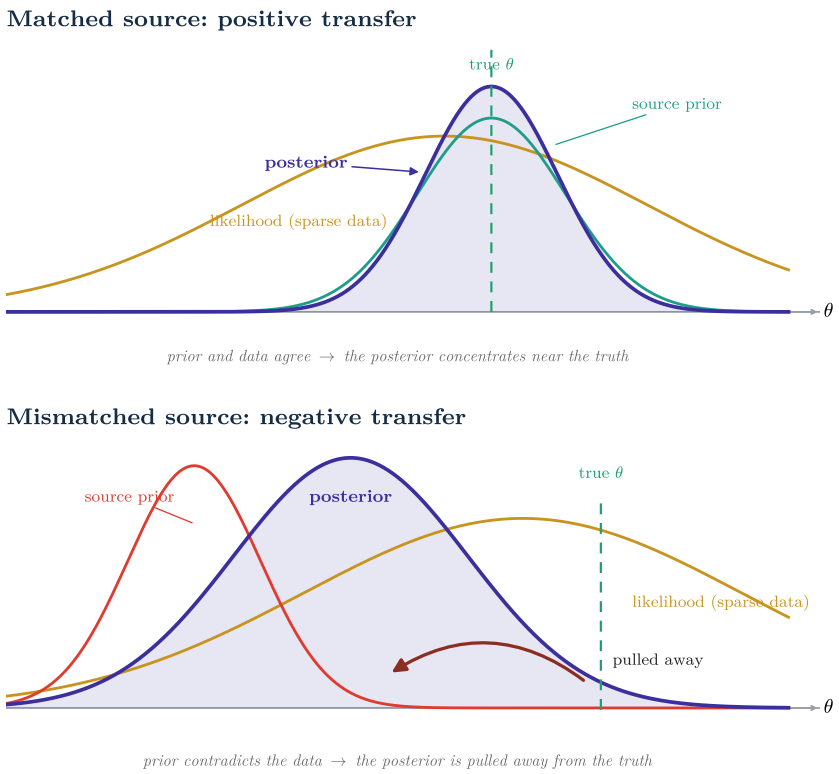


Figure 1-2: Bayesian transfer learning constructs an informative prior from source-domain knowledge, pulling the posterior toward the source estimate. When the source and target domains are well-matched this improves estimation, but when they differ the posterior is pulled away from the true target parameters. Adapted from [1].

The practical difficulty is that the degree of mismatch between source and target is rarely known in advance. Source selection methods that rely on Kullback Leibler (KL) divergence between estimated CPTs are vulnerable to confirmation bias: the source most likely to be selected is the one that most closely matches the current noisy target estimate, which may not be the most informative source. In settings with very limited target data, this bias can be severe.

These risks motivate a more cautious and resilient approach to prior transfer: rather than selecting and applying a source prior directly, it is preferable to first evaluate whether the source prior is compatible with what the target data supports, and to use this compatibility assessment to guide how much influence the source is permitted to exert.

1-3 Problem description and gaps

Existing work has addressed constrained Bayesian inference, prior elicitation, and transfer learning for BBNs as separate problems. Constrained optimization approaches such as Constrained Maximum A Posteriori (CMAP) [15] and Constrained Adjusted Maximum A Posteriori (CaMAP) [16] enforce qualitative constraints during parameter estimation but produce point estimates that do not represent remaining parameter uncertainty, and do not formulate prior construction as an optimization problem over a feasible parameter region. Transfer learning approaches such as MPL-TC [8] and earlier parameter aggregation methods [17] use information from related source domains to improve learning when target data are scarce, but require suitable source-domain information to be available, remain vulnerable to negative transfer when source and target domains differ, and offer no mechanism for evaluating whether a source prior is compatible with target-side evidence before transfer occurs. Credal network approaches [18]

represent parameter uncertainty as sets but provide no mechanism for constructing structured priors within those sets or optimizing a prior with respect to a feasible parameter region.

This thesis addresses these gaps by developing a unified framework for prior construction and evaluation in BBNs under data scarcity. The framework is built on a feasible region that characterizes which parameter configurations remain plausible given the available evidence. The framework operates in two modes: one that constructs an optimized prior from target data and domain constraints alone, and one that evaluates the compatibility of external source priors with the target feasible region before transfer. Both modes are grounded in the same compatibility metrics, providing a coherent and interpretable foundation for prior selection and optimization across different data availability settings.

1-4 Thesis objective and research questions

Based on the problem statement and gaps identified in the literature, the following research objective was formulated.

Research Objective

"The objective of this thesis is to determine whether target-side uncertainty and qualitative domain knowledge can be transformed into informative Bayesian priors that improve parameter learning and inference in BBNs under data scarcity, without relying on source-domain transfer."

To address this objective in a structured way, three research questions are formulated. These research questions are divided into sub-questions to allow for a systematic approach in answering the main questions.

Research Question 1

"How can target-side uncertainty and qualitative knowledge be combined into a coherent feasible parameter region for CPT learning under sparse data?"

1A: How can data-driven intervals be constructed from limited target data?

1B: How can cross-row ordering constraints be incorporated into this representation?

1C: How do these components jointly define a feasible parameter region?

Research Question 2

"How can the compatibility of candidate prior knowledge with the feasible region be quantified and interpreted?"

2A: How can compatibility be measured in terms of prior mass in the feasible parameter space and constraint violation?

2B: How does joint feasible-region evaluation compare to row-wise assessment?

2C: How can incompatibility be decomposed to identify the location and type of inconsistency?

Research Question 3

"How can compatibility diagnostics be used to construct or refine priors under data scarcity, and what is the impact on downstream BBN inference?"

3A: How can the compatibility metrics be integrated into a cost function for prior construction?

3B: How do optimized priors compare to uniform baselines and transfer learning benchmarks in terms of CPT estimation accuracy?

3C: Does the improvement in CPT accuracy propagate to improved hypothesis inference in the HMF?

1-5 Contribution and thesis outline

The main contributions of this thesis are the following.

1. **A feasible-region formulation for constraint-informed prior construction.** Target-side uncertainty is represented as a feasible parameter region derived from Wilson score confidence intervals and domain-informed cross-row constraints. This provides a structured representation of admissible CPT parameter configurations under data scarcity without requiring a fully specified prior distribution.
2. **Compatibility metrics for prior feasible-region evaluation.** Two complementary compatibility metrics are introduced to quantify how well a candidate prior aligns with the feasible region: prior mass, measuring the probability assigned to admissible parameter configurations, and expected violation, quantifying the degree of infeasibility associated with prior samples. Together, these metrics provide a principled basis for evaluating candidate prior knowledge before transfer or optimization.
3. **A violation decomposition for actionable incompatibility analysis.** Global incompatibility between a candidate prior and the feasible region is decomposed into row-wise and constraint-wise contributions, enabling targeted diagnosis of the location and type of inconsistency rather than treating incompatibility as a single aggregate quantity.
4. **A feasible-region-based prior optimization framework.** The compatibility metrics are incorporated into a composite cost function over a parametric Beta prior family, combining feasible-region coverage, constraint satisfaction, centrality within the feasible region, and target-uncertainty regularization. The resulting framework constructs informative priors directly from target-side information without requiring source-domain data.
5. **Empirical validation through parameter learning and network inference.** The proposed framework is evaluated through controlled simulation experiments and compared with uniform priors and MPL-TC transfer-learning baselines. The results demonstrate consistent improvements in CPT estimation accuracy under data scarcity and show that these gains propagate to improved network-level inference within the HMF setting.

Outline and roadmap

This thesis is organized as follows. Chapter 2 introduces the theoretical and methodological background. It covers BBN parameter learning under data scarcity, constrained parameter-learning methods, transfer learning for BBNs including the MPL-C and MPL-TC frameworks,

and imprecise probability theory as a robustness framework. The chapter concludes by positioning the present work relative to existing approaches and identifying the specific gaps addressed by this thesis.

Chapter 3 addresses the methodological framework of this thesis. It introduces the feasible-region construction from Wilson intervals and domain-informed constraints, the compatibility metrics used for prior evaluation, the violation decomposition and actionable violation analysis, the prior optimization cost function and regularization strategy, and the optimization procedure. It also describes the network-inference procedure using optimized priors.

Chapter 4 presents the experimental evaluation. It covers four experimental components: validation of the compatibility metrics, evaluation of actionable violation analysis, prior optimization and parameter learning including a benchmark comparison with MPL-TC, and network-level HMF inference evaluation.

Chapter 5 summarizes the conclusions and main contributions of the thesis. It provides answers to each of the three research questions, summarizes the main contributions, discusses the limitations of the proposed framework, along with recommendations for future work.

Chapter 2

Background and related work

This chapter reviews the theoretical and methodological background relevant to parameter learning in discrete BBNs under data scarcity. It introduces the formal structure of BBNs and standard parameter-learning methods, discusses their limitations when data is limited, and motivates the use of qualitative constraints. It then describes the MPL-TC transfer-learning framework used as a benchmark, reviews how CPT uncertainty propagates to downstream inference queries, and introduces the HMF as both the application context and the specific network model used in this thesis. The chapter concludes by identifying the gaps in existing work that motivate the feasible-region-based framework developed in Chapter 3.

2-1 Bayesian Belief Networks

BBNs are frameworks for probabilistic reasoning under uncertainty in which the nodes represent random variables and the directed edges encode direct causal or dependency relationships between variables [19, 13]. A BBN is a structured probabilistic model that combines a graphical representation of conditional independence relations with local probability tables, enabling compact representation of joint distributions over multiple variables. BBNs are widely used for reasoning under uncertainty in domains where causal structure can be encoded as a directed graph [20, 5, 21]. This section introduces the formal foundations relevant to this thesis. Section 2-1-1 defines the graph structure and CPT parameterization. Section 2-1-2 describes ML estimation as the standard approach to parameter learning from data.

2-1-1 Definition and parameterization

A BBN is a structured probabilistic model defined by a Directed Acyclic Graph (DAG) together with a collection of local CPTs. The DAG encodes conditional independence relations that permit the factorization of the joint distribution, while each CPT specifies the conditional probabilities associated with each node in the model. A BBN is defined as (G, Θ), where $G = (\mathbf{X}, E)$ is a directed acyclic graph with variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ and directed edges E, and Θ collects the CPT parameters [22]. Each node corresponds to a discrete random variable and each directed edge $X_i \to X_j$ encodes a direct probabilistic dependence of X_j on X_i.

For any ordering $X_1, \ldots, X_n$, the graph encodes the Markov condition: each variable is conditionally independent of its non-descendants given its parents [20, 23, 24].

The CPT of node X_i is parameterized by entries

$$\theta_{ij}^{k} = \mathbb{P}(X_i = x_k \mid \Pi_i = x_j), \tag{2-1}$$

where $k \in \{1, \ldots, r_i\}$ indexes the states of X_i, $j \in \{1, \ldots, s_i\}$ indexes the joint parent configurations, and Π_i denotes the parent set of X_i. For each parent configuration j, parameters must satisfy:

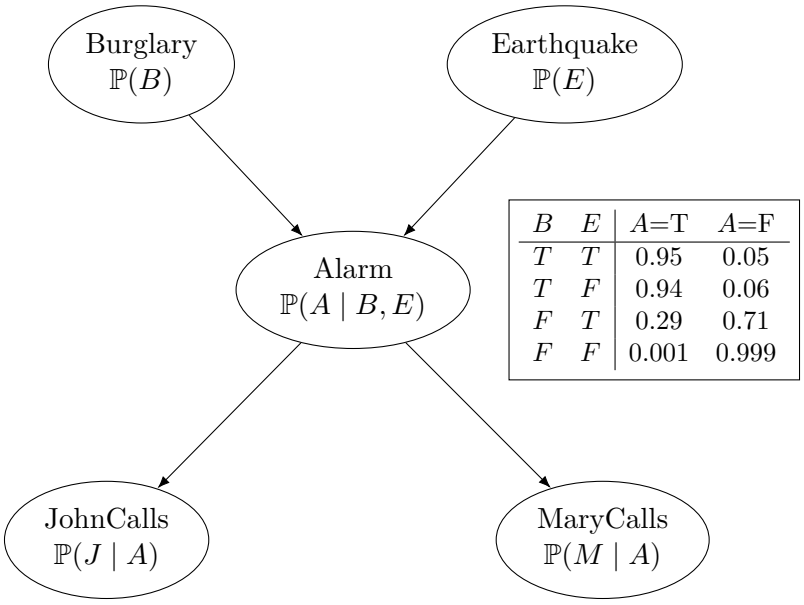
$$\sum_{k=1}^{r_i} \theta_{ij}^{k} = 1, \quad \theta_{ij}^{k} \geq 0, \quad \forall j. \tag{2-2}$$

If a variable X_i has no parents, its CPT reduces to its marginal distribution $\mathbb{P}(X_i)$.

The joint distribution factorizes according to the parent sets:

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid \Pi_i). \tag{2-3}$$

This factorization, illustrated in Figure 2-1 for the classic burglary-alarm network, allows the joint distribution over n variables to be represented compactly through local conditional probability tables rather than requiring an explicit specification of the full joint distribution [25].



B	E	A=T	A=F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

Figure 2-1: The burglary-alarm BBN. Each node is associated with a local CPT. The row (B=T, E=F) specifies $\mathbb{P}(A$=T $\mid B$=T, E=F$) = 0.94$. JohnCalls and MaryCalls are conditionally independent of Burglary and Earthquake given the Alarm state.

2-1-2 Maximum likelihood estimation

Given a fixed network structure G and a dataset $\mathbf{D} = \{x^{(1)}, \ldots, x^{(m)}\}$ of complete observations, the goal of parameter learning is to estimate $\Theta = \{\theta_{ijk}\}$ from data. Let N_{ij}^{k} denote the number of observations in which $X_i = x_k$ while its parents take configuration j, and let $N_{ij} = \sum_k N_{ij}^{k}$. The ML estimate is:

$$\hat{\theta}_{ijk}^{\mathrm{ML}} = \frac{N_{ij}^{k}}{N_{ij}}, \qquad N_{ij} > 0. \tag{2-4}$$

Each CPT row is obtained by normalizing the observed counts for that parent configuration. As a simple illustration, consider the Alarm node in Figure 2-1. If the parent configuration

$(B = \mathrm{T}, E = \mathrm{T})$ is observed 20 times with 19 occurrences of $A = \mathrm{T}$, the ML estimate is:

$$\hat{\mathbb{P}}(A = \mathrm{T} \mid B = \mathrm{T}, E = \mathrm{T}) = \frac{19}{20} = 0.95,$$

matching the CPT entry shown in the figure. This estimator is consistent: as the dataset grows, $\hat{\theta}^{\mathrm{ML}}$ converges to the true parameters under standard regularity conditions [26, 27]. However, this guarantee provides little practical protection when observed counts are small, as examined in Section 2-2.

2-2 Parameter learning under data scarcity

When training data is limited, standard parameter-learning methods produce estimates that are unstable, poorly calibrated, or practically unspecifiable, undermining the reliability of downstream inference. This section examines these limitations and reviews the approaches proposed in the literature to address them.

2-2-1 Limitations of ML estimation

ML estimation exhibits three serious limitations when data is scarce [28]. First, zero counts produce zero probability estimates: if $N_{ij}^k = 0$ for some state, then $\hat{\theta}_{ijk}^{\mathrm{ML}} = 0$, making that outcome appear impossible regardless of its true probability. Second, when parent configurations are rare, estimates become highly variable, reflecting sampling noise rather than the underlying distribution. Third, in networks with many parent configurations, most configurations may never be observed in a limited dataset, leaving large portions of the CPT undefined [22]. These effects propagate to downstream inference: small changes in a limited dataset can produce large shifts in posterior conclusions [29].

2-2-2 Bayesian estimation and prior specification

In contrast to ML, which uses only the data, Bayesian estimators, including MAP and posterior mean, incorporate prior information and are therefore less prone to overfitting when data is scarce [29, 30, 31, 32]. Bayesian parameter estimation treats CPT entries as random variables and updates beliefs about their values by combining a prior distribution with the observed data likelihood through Bayes' rule:

$$\mathbb{P}(\theta \mid \mathbf{D}) \propto \mathbb{P}(\mathbf{D} \mid \theta)\, \mathbb{P}(\theta), \tag{2-5}$$

where $\mathbb{P}(\theta)$ encodes prior beliefs about the parameters before observing data and $\mathbb{P}(\mathbf{D} \mid \theta)$ is the likelihood of the observed counts under those parameters [29, 10, 30]. When observations are abundant, the likelihood dominates and the posterior concentrates near the ML estimate regardless of the prior; when observations are scarce, the prior exerts greater influence and its specification becomes consequential.

Conjugate Beta-Binomial updating

For a binary node X_i, each CPT row has one free parameter $\theta_{ij} = \mathbb{P}(X_i = 1 \mid \Pi_i = j)$. The natural conjugate prior for this parameter is the Beta distribution [33, 26, 34]:

$$\theta_{ij} \sim \mathrm{Beta}(\alpha_{ij}, \beta_{ij}), \tag{2-6}$$

parameterized by its mean $\mu_{ij} = \alpha_{ij}/(\alpha_{ij} + \beta_{ij})$ and concentration $\tau_{ij} = \alpha_{ij} + \beta_{ij}$. The mean encodes the expected parameter value while the concentration acts as an equivalent sample size: it determines how strongly the prior resists updating when new observations arrive. Conjugacy ensures that after observing counts (N^1_{ij}, N^0_{ij}), where $N_{ij} = N^1_{ij} + N^0_{ij}$, the posterior is again Beta:

$$\theta_{ij} \mid \mathbf{D} \sim \text{Beta}(\alpha_{ij} + N^1_{ij},\ \beta_{ij} + N^0_{ij}), \tag{2-7}$$

with posterior mean:

$$\hat{\theta}^{\text{post}}_{ij} = \frac{\alpha_{ij} + N^1_{ij}}{\tau_{ij} + N_{ij}}. \tag{2-8}$$

This expression interpolates between the prior mean μ_{ij} and the ML estimate N^1_{ij}/N_{ij}, with the balance governed by the ratio of prior concentration τ_{ij} to observed count N_{ij}. The posterior variance is:

$$\text{Var}(\theta_{ij} \mid \mathbf{D}) = \frac{\hat{\theta}^{\text{post}}_{ij}(1 - \hat{\theta}^{\text{post}}_{ij})}{\tau_{ij} + N_{ij} + 1}, \tag{2-9}$$

showing that uncertainty decreases as either more data is observed or prior concentration increases. The MAP estimate is:

$$\hat{\theta}^{\text{MAP}}_{ij} = \frac{N^1_{ij} + \alpha_{ij} - 1}{N_{ij} + \tau_{ij} - 2}. \tag{2-10}$$

The multivariate generalization to nodes with more than two states replaces the Beta prior with a Dirichlet distribution and Binomial counts with Multinomial counts, preserving conjugacy [20, 35, 36]. In the general case, the posterior remains Dirichlet with updated hyperparameters $\alpha_{ijk} + N_{ijk}$, and the MAP estimate takes the standard closed form [32, 37, 26].

Prior specification in practice

The performance of Bayesian estimation depends on the choice of prior hyperparameters $(\alpha_{ij}, \beta_{ij})$. Two broad strategies appear in the BBN literature [26].

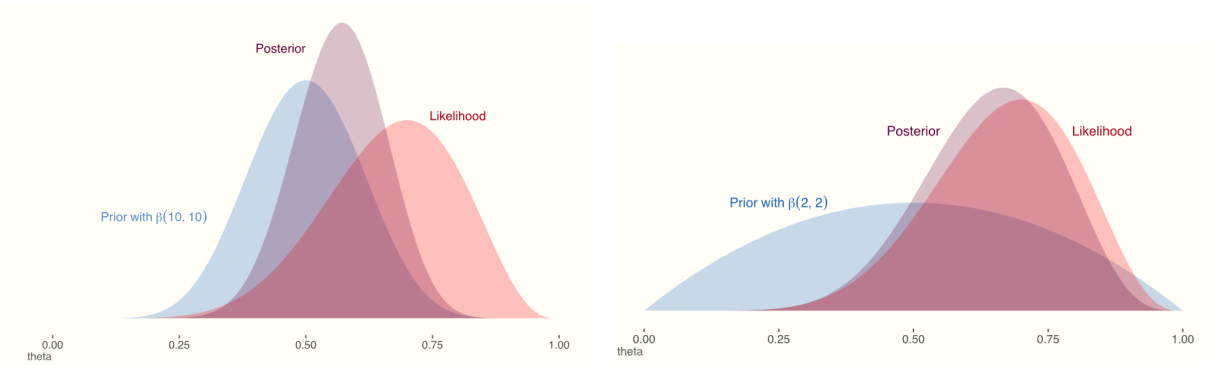
An *uninformative prior* expresses the absence of prior knowledge [38]. The standard choice is the uniform prior $\text{Beta}(1, 1)$, which contributes one pseudo-observation to each outcome and assigns equal probability to all parameter values. Alternative uninformative constructions exist, such as the Jeffreys prior $\text{Beta}(0.5, 0.5)$, which is invariant under reparameterization [39], but the practical difference between these choices is rarely consequential for BBN parameter learning.

An *informative prior* encodes substantive domain knowledge through the prior mean μ_{ij} and concentration τ_{ij}. The concentration is often elicited through its equivalent sample size interpretation: a practitioner specifies how many real observations their prior belief is worth [26, 38]. Figure 2-2 illustrates how these two strategies respond differently to the same observed data: an informative prior resists updating and keeps the posterior near its prior mean, while a weak prior shifts readily toward the likelihood.

2-2-3 Limitations of prior-based estimation

Dirichlet priors stabilize estimation by introducing pseudo-counts, but introduce their own challenges under data scarcity [16]. Choosing the concentration τ_{ij} is non-trivial: a large equivalent sample size may dominate the likelihood and make posteriors insensitive to evidence, while a small value provides insufficient regularization when counts are near zero. More fundamentally, experts generally cannot reliably specify full Dirichlet hyperparameter vectors [16, 40]. They are typically more comfortable providing comparative or qualitative judgments [41], such as:

$$\mathbb{P}(I = \text{T} \mid H = \text{T}) > \mathbb{P}(I = \text{T} \mid H = \text{F}),$$



(a) Informative prior: Beta(10, 10) **(b)** Uninformative prior: Beta(2, 2)

Figure 2-2: Effect of prior concentration on posterior updating given the same observed data. A high-concentration prior resists updating and keeps the posterior near the prior mean; a low-concentration prior shifts readily toward the likelihood [2].

which constrain relationships between parameters but do not correspond to any natural Dirichlet prior. In intelligence-analysis settings such as the HMF, this elicitation gap is particularly pronounced: analysts reason in terms of comparative likelihoods and causal orderings, not numerical pseudo-counts [42].

These limitations, zero estimates from ML, high variance under sparse counts, and the elicitation gap for prior hyperparameters, establish the need for approaches that bound plausible parameter values directly from data and encode expert knowledge through qualitative comparisons rather than numerical pseudo-counts. Both needs are addressed in Chapter 3.

2-2-4 Interval estimation under data scarcity

A natural response to estimation uncertainty under small samples is to replace point estimates with confidence intervals that characterize the range of plausible parameter values. For a binary node with observed counts (N_{ij}^1, N_{ij}^0), a $(1-\alpha)$ confidence interval for θ_{ij} can be constructed using the Wilson score method [43]:

$$\left[\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}\right], \tag{2-11}$$

where $\hat{p} = N_{ij}^1/N_{ij}$ is the sample proportion, $n = N_{ij}$ is the total count for that row, and z is the $(1-\alpha/2)$ quantile of the standard normal distribution. Unlike the Wald interval, which centers directly on $\hat{p}$ and collapses to a degenerate point when $\hat{p} \in \{0, 1\}$, the Wilson interval centers on the shrinkage estimate

$$\tilde{p} = \frac{\hat{p} + z^2/2n}{1 + z^2/n}, \tag{2-12}$$

which is always bounded away from the extremes, ensuring positive interval width even under boundary counts and better empirical coverage at small sample sizes [44, 45]. In the limit $n \to 0$, the interval approaches $[0, 1]$, consistent with the complete uncertainty of the parameters.

Each interval defines a range of parameter values consistent with the observed data at a specified confidence level, producing data-driven bounds without requiring expert elicitation of prior hyperparameters. When applied to each CPT row independently, these intervals define a hyper-rectangular region in parameter space whose width contracts as sample size grows, reflecting increasing data informativeness. The construction of such intervals and their role in defining a feasible parameter region is developed in Chapter 3.

2-3 Qualitative constraints and constrained parameter learning

When numerical prior specification is infeasible, qualitative domain knowledge can still be incorporated into BBN parameter learning through constraints on admissible CPT entries [46]. Rather than requiring experts to assign pseudo-counts or probability values, constraints capture comparative judgments, such as ordering relationships between parameters, as linear inequalities that restrict the parameter space. This section reviews how such constraints are formalized, used in the literature, and how they reduce the admissible parameter space geometrically.

2-3-1 The role of qualitative constraints

Constraints provide a principled way to incorporate expert knowledge without requiring precise probability estimates [47, 48, 49, 50, 51]. The constraints most directly relevant to this thesis are cross-row ordering constraints, also referred to in the literature as exterior constraints, which express that a CPT entry under one parent configuration should be ordered relative to the same entry under a different parent configuration [52, 48]. This is in contrast to interior constraints, which restrict relationships between entries within the same parent configuration column [53, 54]. Formally, for two parent configurations j_1 and j_2 of node X_i:

$$\theta^k_{ij_1} \geq \theta^k_{ij_2}. \tag{2-13}$$

As a concrete example, consider the HMF model in Figure 2-6, in which two hypotheses influence an indicator I. An analyst may know that the indicator is more likely to be present when the first hypothesis holds than when neither does, without being able to assign precise probabilities to either case. This comparative judgment translates directly into a cross-row ordering constraint:

$$\mathbb{P}(I = \mathrm{T} \mid H_1 = \mathrm{T}, H_2 = \mathrm{F}) \geq \mathbb{P}(I = \mathrm{T} \mid H_1 = \mathrm{F}, H_2 = \mathrm{F}), \tag{2-14}$$

which rules out all CPT parameterizations in which the indicator is equally or less likely when H_1 holds than when neither hypothesis holds. Such constraints are linear inequalities over CPT entries and define a convex feasible parameter region. The theoretical foundation for incorporating linear inequality constraints into BBN parameter learning is established in [55], showing that the resulting constrained optimization satisfies Karush–Kuhn–Tucker conditions and admits a unique global optimum when the feasible set is convex.

Cross-row ordering constraints are a special case of the broader class of cross-CPT exterior constraints studied in the constrained learning literature [48, 52, 56, 15, 16]. It has been shown empirically that such qualitative sign constraints improve estimation accuracy at small sample sizes, since they eliminate parameterizations inconsistent with known causal relationships [57]. Monotonicity constraints have further been shown to reduce estimation variance substantially when data is scarce, and this benefit persists even under moderate misspecification of the constraint direction [56].

Approximate equality constraints, also referred to as closeness constraints, require two CPT entries to differ by at most ε:

$$|\theta^k_{ij_1} - \theta^k_{ij_2}| \leq \varepsilon, \tag{2-15}$$

encoding the belief that two parent configurations have a similar but not necessarily equal effect on the child node [48, 49, 56]. Unlike ordering constraints, which arise naturally from expert causal reasoning, closeness constraints are more commonly used as a technical regularization device to prevent CPT rows from diverging beyond what the domain supports, rather than as direct translations of analyst judgment.

2-3-2 Constrained learning methods

The literature has developed two broad classes of constrained parameter-learning methods.

Constraint-based Bayesian methods combine qualitative constraints with an explicit Dirichlet prior, preserving conjugate Bayesian updating while restricting admissible parameters to a feasible region [50]. Representative approaches include Constrained Bayesian Estimation (CBE) [22], Constrained Maximum A Posteriori (CMAP) [15], Constrained Adjusted Maximum A Posteriori (CaMAP) [16] and Qualitative Maximum A Posteriori (QMAP) [58].

Prior-distribution-free methods rely solely on constraints to guide estimation without specifying a prior. Examples include isotonic regression for monotonic influences [57], convex-constrained ML estimation [48], and evolutionary constrained ML algorithms [28]. These approaches optimize the likelihood subject to feasibility conditions and remain non-Bayesian in the statistical sense. They avoid the need to specify prior strength, an advantage when experts cannot reliably quantify their beliefs numerically.

Both classes return a single optimal CPT estimate, leaving residual parameter uncertainty unrepresented. This limitation motivates approaches that construct a prior distribution over a constrained parameter region rather than returning a point estimate, as discussed in Chapter 3.

A common approach to solving constrained parameter-learning problems is to convert the constrained formulation into an unconstrained one using penalty methods [59, 60]. Rather than enforcing constraints directly, the constraint violations are added to the data log-likelihood as penalty terms, so that a single objective combines fit to the data with a penalty that grows in proportion to the degree of constraint violation [60]. This transformation allows standard unconstrained optimization algorithms to be applied while still discouraging solutions that violate the imposed constraints.

2-3-3 Parameter spaces under constraints

Adding constraints reduces the admissible parameter space [22, 3, 49]. Cross-row ordering constraints like Eq. 2-13 couple parameters from different CPT rows simultaneously, producing a convex feasible region whose boundary is not axis-aligned. This contrasts with interval constraints alone, which yield an axis-aligned box. The combined effect is illustrated geometrically in [3], where the constraint $P_1 > P_2$ reduces the unit square to a triangular convex subset; intersecting this with interval bounds yields a convex polytope. Linear inequality constraints on CPT parameters preserve convexity within the parameter simplex; [55] establishes this for both within-column and cross-row (cross-parent) inequality constraints, and the property is exploited throughout the constrained-learning literature [48, 15].

This construction is related to the credal set framework of imprecise probability theory [18, 61], in which a convex set of admissible distributions represents epistemic uncertainty about parameters. A credal network generalizes a BBN by associating each node with such a set rather than a single CPT, with inference returning bounds on query probabilities rather than point values. Within this framework the admissible set is itself the object of inference: uncertainty is propagated by optimizing over the set. The framework does not, however, address the construction of a structured prior distribution within the set, nor the evaluation of an externally supplied prior against it. Geometric size measures such as volume characterize the extent of such a set but likewise say nothing about how a given prior distribution is positioned relative to it [62].

These properties of constraint-informed parameter spaces form the geometric foundation for the feasible-region construction developed in Chapter 3.

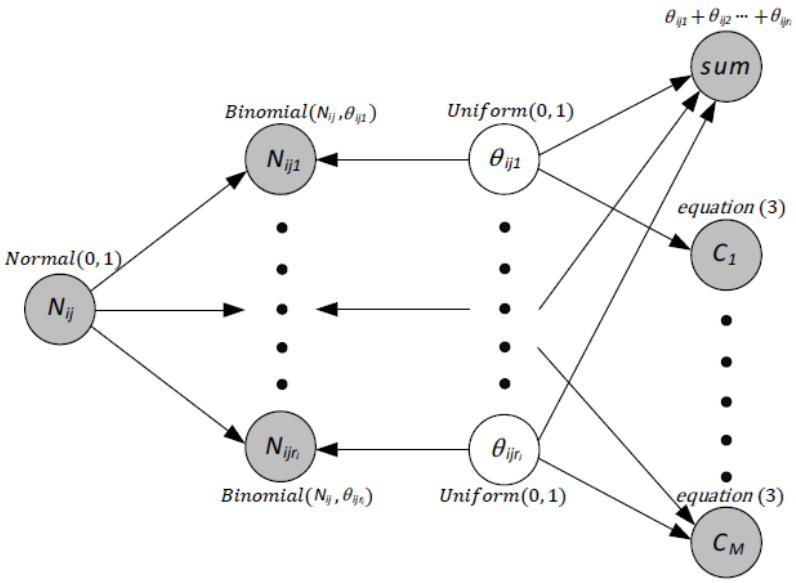


Figure 2-3: Graphical model representation of MPL-C with M constraints [3].

2-4 Transfer learning for BBN parameter learning: MPL-TC

When target-domain data is scarce, one natural strategy is to exploit knowledge from a related source domain. The Multinomial Parameter Learning with Transferred priors and Constraints (MPL-TC) framework [8] formalizes this idea for BBN parameter learning. The method assumes access to a source dataset generated from a domain related to but distinct from the target, for example a different time period, geographic region, or conflict context in HMF applications.

The MPL-TC framework is the most recent in a line of related models. It builds on MPL-C [3] (Figure 2-3), which combines target data with expert constraints under a uniform prior over the constrained region, and on an extension of MPL-C that incorporates exterior (cross-CPT) constraints through an auxiliary BBN [52]. MPL-TC extends this line to the transfer setting, replacing the uniform prior with an informative prior derived from a related source domain.

MPL-TC proceeds in two stages. In the first stage, bootstrap samples are drawn from the source dataset and ML estimates are computed for each bootstrap replicate. The variability of these estimates across replicates is used to construct an informative prior distribution over target CPT parameters, capturing both the location and the uncertainty of the source estimate. MPL-TC replace the uniform prior in MPL-C with a Truncated Normal prior[8]:

$$\zeta^s_{i'jk} = \text{TNormal}(\mu_{ijk}, \sigma_{ijk}, 0, 1).$$

In the second stage, the resulting source-derived prior is updated conjugately with the target inference data.

A practical limitation of MPL-TC is its dependence on source domain quality. The method assumes that at least one source domain sampled from a distribution similar to the target is available. In operational settings, however, the degree of mismatch between source and target is typically unknown, and source selection based on scarce target data may itself produce confirmation bias toward sources that superficially resemble the current noisy target estimate rather than the true underlying distribution [8]. When source and target are not sampled from similar distributions, directly applying parameters from the source domain may result in negative transfer [1, 8]. Figure 1-2 contrasts this failure mode with the favorable case: a well-matched source prior pulls the posterior toward the truth, whereas a mismatched source prior pulls it away.

MPL-TC incorporates a compatibility screening step before transfer: a source fragment is excluded if its structure differs from the target fragment or if its parameters fall outside the target constraint ranges $\Omega_{C_i^t}$ [8]. This screening has two limitations. First, the constraints used are within-column range constraints; cross-row ordering constraints spanning multiple parent configurations of the same node are not evaluated. Each CPT column is assessed independently, so a source prior may pass screening while still violating ordering relationships across rows. The paper itself acknowledges that incorporating cross-node and cross-domain constraints to guide transfer remains an open direction [8]. Second, compatibility is treated as a binary condition: a source fragment is either excluded or included, with no graded measure of how severely it violates the target constraints or how compatible it is with the feasible parameter region. Neither limitation is addressed within the MPL-TC framework.

This limitation is particularly relevant when qualitative knowledge is expressed through cross-row monotonicity constraints, since compatibility can no longer be assessed independently for each CPT row. Although MPL-TC incorporates compatibility screening before transfer, compatibility is assessed at the level of source fragments and local parameter constraints. The method does not provide an explicit feasible-region-based compatibility framework that evaluates candidate source priors against a joint target uncertainty representation and cross-row constraint structure prior to transfer. MPL-TC serves as the primary transfer learning benchmark in Chapter 4.

2-5 CPT uncertainty propagation in BBNs

When BBN parameters are treated as random variables rather than fixed values, uncertainty in the CPT parameters propagates to uncertainty in query responses, which are themselves random variables [63]. This perspective is referred to as second-order probabilistic inference [9]: rather than returning a single point estimate for a query such as $P(H \mid E)$, the network returns a distribution over plausible query values induced by parameter uncertainty.

Analytical expressions for the mean and asymptotic variance of this query response distribution can be derived using a delta-method approximation, showing it is asymptotically normal [63]. However, at small sample sizes the normal approximation may be inaccurate, as query distributions are bounded to $[0, 1]$ and can exhibit skewness. A non-parametric alternative is to characterize the query distribution by Monte Carlo sampling from the parameter posteriors and applying Kernel Density Estimation (KDE) to the resulting samples [64], which imposes no distributional assumption and naturally handles boundary effects. This approach is used in Section 4-4-3 to characterize posterior query uncertainty under different prior construction strategies.

This motivates evaluation at both the CPT level and the network query level, as formalized in the evaluation framework described in Chapter 3.

2-6 The Hypothesis Management Framework

The HMF is the operational context for this thesis. It is a design pattern for BBNs intended for intelligence domains where causal structure is contested or only partially known, designed for modularity and ease of maintenance [65, 66].

The central aim is modularity and usability without sacrificing decision-support reliability: analysts should obtain coherent posterior assessments given the same evidence, while domain experts should face a minimal elicitation burden. HMF imposes a role-based topology with three node types:

- A **hypothesis** (H) is a potential explanation or conclusion based on the observations. The state of a hypothesis is unknown as the hypothesis cannot be observed directly. Usually a model consists of multiple hypotheses.
- An **indicator** (I) is an observable phenomenon with diagnostic value for one or more hypotheses. Indicators are typically modelled as discrete random variables with states corresponding to qualitatively different conditions (e.g., present/absent, low/medium/high activity, detected/not detected). Indicators mediate between abstract hypotheses and concrete observations.
- A **statement** (S) is a single unique observation tied to a clearly specified source, confirming or refuting at least one indicator. With statements, knowledge about the reliability of observations can be separated from knowledge about the occurrence of the indicators. This can be any kind of intelligence report.

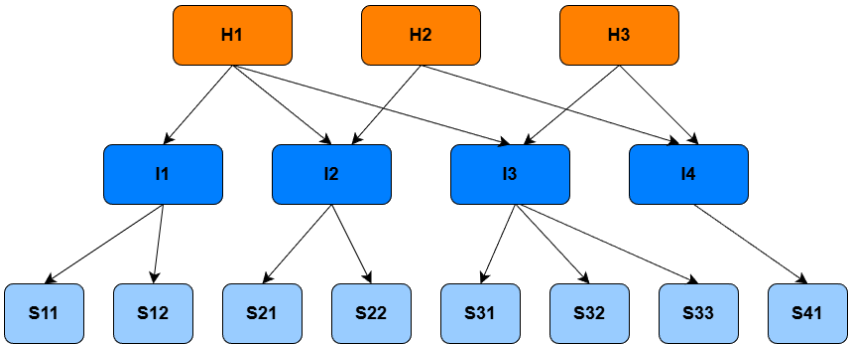


Figure 2-4: A non-disjoint HMF model. Hypotheses, indicators, and statements form three distinct layers with directed edges encoding diagnostic relationships.

Modeling assumptions

The following assumptions hold in the standard HMF [65]:

- **Independent hypotheses:** competing hypotheses are modeled as marginally independent, $\mathbb{P}(H_1, H_2, H_3) = \mathbb{P}(H_1)\,\mathbb{P}(H_2)\,\mathbb{P}(H_3)$, so direct dependencies between hypotheses are not representable.
- **Independent indicators:** direct dependencies between indicators are not modeled, even when indicators may be correlated in reality.
- **Independent statements:** direct dependencies between statements are not modeled, which limits representation of shared source biases or correlated reporting mechanisms.
- **No synergistic effects** *(standard HMF):* in the default non-disjoint parameterization, joint influences of multiple hypotheses on a single indicator are suppressed; only marginal relations $\mathbb{P}(I \mid H_k)$ are specified per hypothesis.

The fourth item reflects a deliberate design choice in the standard HMF: allowing synergy requires a multi-parent CPT whose number of rows grows exponentially with the number of hypotheses, increasing elicitation complexity and the risk of unstable posteriors under sparse data. For operational models with many hypotheses this restriction is essential for tractability and usability.

This thesis relaxes this fourth assumption: the experimental model in Chapter 4 uses a two-hypothesis indicator with a full joint CPT $\mathbb{P}(I \mid H_1, H_2)$ rather than independent marginals, so that synergistic effects between hypotheses can be represented. At this scale the joint table has only four rows, so the exponential-scaling concern that motivates the standard restriction does not apply.

HMF models are categorized as *disjoint* or *non-disjoint* [4]. In a disjoint model, hypotheses form a full partition of the hypothesis space so that exactly one must be true; prior and posterior probabilities must sum to one. In a non-disjoint model, multiple hypotheses may hold simultaneously and probabilities need not sum to one. The present thesis uses a non-disjoint HMF with two binary hypotheses, one indicator node with joint CPT $\mathbb{P}(I \mid H_1, H_2)$, and one statement node.

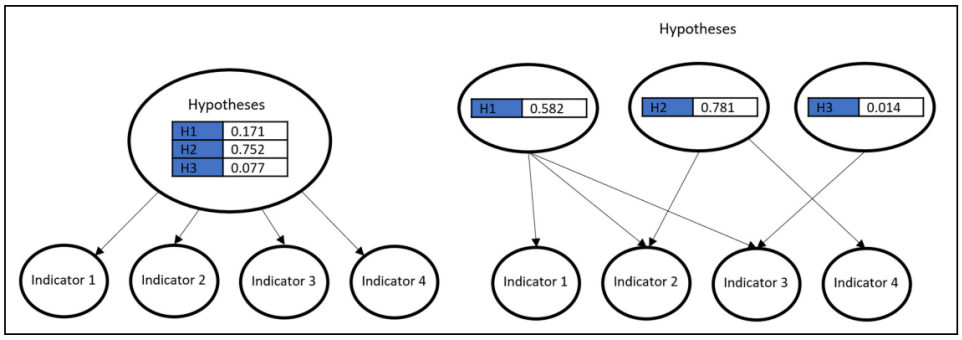


Figure 2-5: Schematic representation of a disjoint model (left) and a non-disjoint model (right) [4].

Inference in the HMF

The primary operational inference mode of the HMF is diagnostic: evidence observed at the statement layer propagates upward to revise hypothesis beliefs [66]. Formally, given evidence $E = e$ and query variables Q, inference seeks the posterior:

$$\mathbb{P}(Q \mid E = e) = \frac{\mathbb{P}(E = e \mid Q)\,\mathbb{P}(Q)}{\mathbb{P}(E = e)}. \tag{2-16}$$

Given a statement $S = s$, the posterior over the indicator is updated first:

$$\mathbb{P}(I \mid S = s) \propto \mathbb{P}(S = s \mid I)\,\mathbb{P}(I),$$

where $\mathbb{P}(I)$ is the prior predictive distribution over the indicator induced by the hypothesis priors and the indicator CPT. This belief then propagates upward to revise hypothesis posteriors. In the non-disjoint model used in this thesis, the indicator is jointly conditioned on both hypotheses, so the update takes the form:

$$\mathbb{P}(H_1, H_2 \mid S = s) \;\propto\; \mathbb{P}(H_1)\,\mathbb{P}(H_2) \sum_i \mathbb{P}(I = i \mid H_1, H_2)\;\mathbb{P}(I = i \mid S = s). \tag{2-17}$$

Here $\mathbb{P}(I \mid H_1, H_2)$ is a full joint CPT with four rows, one per binary parent configuration, not a product of independent marginals. Figure 2-6 illustrates this structure for the minimal HMF used throughout this thesis.

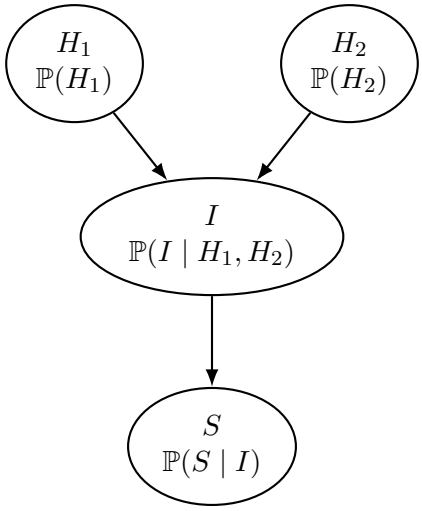


Figure 2-6: Small HMF network with two hypotheses, one indicator, and one statement.

Given the layered structure and the first three independence assumptions of the HMF, exact inference is tractable via standard belief propagation [25, 4]: the sparsely connected topology keeps clique sizes small during moralization, so junction-tree or variable elimination methods scale efficiently even for operationally sized models. Parameter uncertainty in the joint CPT $\mathbb{P}(I \mid H_1, H_2)$ propagates non-linearly through this inference chain to the hypothesis posteriors $P(H_k \mid S)$, motivating evaluation at both the parameter and query level [67, 9].

2-7 Summary

This chapter reviewed the background and theoretical foundations of BBNs, parameter learning under data scarcity, qualitative constraint incorporation, transfer learning, and the HMF. The literature shows that data scarcity creates substantial challenges for reliable CPT estimation, motivating the use of Bayesian priors, qualitative constraints, and transfer-learning approaches.

Constrained learning methods such as MPL-C, CMAP, and CaMAP demonstrate that qualitative domain knowledge can improve parameter estimation by restricting the admissible parameter space. However, these methods primarily focus on estimating constrained parameter values and do not address the construction of informative prior distributions from target-side uncertainty and constraints.

Transfer-learning methods such as MPL-TC address data scarcity by transferring information from related source domains. While these methods include compatibility screening procedures, they rely on the availability of suitable source data and provide limited support for evaluating candidate source priors against a jointly defined target feasible region that incorporates both target uncertainty and cross-row qualitative constraints.

Consequently, two gaps emerge from the literature. First, there is currently no framework that constructs informative priors directly from target-side uncertainty and qualitative constraints through an explicit optimization procedure, without requiring source-domain information. Second, there is limited support for systematically evaluating how well candidate source priors align with a target-domain feasible parameter region before transfer occurs.

The next chapter addresses these gaps by introducing a feasible-region-based framework for prior evaluation and optimization. First, target-side uncertainty and qualitative constraints are combined into a feasible parameter region. Second, compatibility metrics are developed to evaluate both source priors and optimized priors with respect to this region. Finally, these

metrics are incorporated into a cost function for constructing informative priors directly from target-side information.

Chapter 3

Methodology

This chapter presents a framework for Bayesian prior optimization in BBNs under data scarcity. The central idea is to construct informative priors from target-side uncertainty and domain-informed constraints, providing an alternative to both non-informative priors and transfer-learning approaches that rely on source-domain information.

The method starts from the construction of a feasible parameter region from target-side uncertainty and qualitative constraints, addressing RQ1 and its subquestions. Compatibility with this region is then quantified through prior mass and expected violation metrics, and further decomposed into row-wise and constraint-wise contributions, addressing RQ2 and subquestions. In addition, a diagnostic violation decomposition is introduced to localize incompatibility within the parameter space, enabling targeted refinement of prior knowledge and supporting analyst interpretation. Finally, these diagnostics are integrated into a cost function for prior construction, addressing RQ3A.

The main contribution of this chapter is a feasible-region-based prior optimization procedure that learns structured priors consistent with target-side information while remaining informative for downstream network-level Bayesian inference. The cost function combines the compatibility metrics with a centering regularization term that discourages priors concentrated near the boundary of the feasible region, and a target-uncertainty regularization term that prevents overconfident priors.

The framework operates in two modes. In its evaluation mode, the compatibility metrics are applied to assess and rank candidate source priors against the target-side feasible region, enabling systematic source selection and identification of parameters driving incompatibility. In its construction mode, the same metrics are incorporated directly into a cost function to construct a structured prior entirely from target-side information, requiring no source-domain data. While the primary focus is on defining the cost function itself, this chapter also presents the optimization procedure used to solve it and motivates the associated design choices.

The remainder of this chapter is organized as follows. Section 3-1 introduces the problem setting. Section 3-2 defines the feasible parameter region. Section 3-3 presents the two compatibility metrics. Section 3-4 applies these metrics to evaluate and rank candidate source priors. Section 3-5 introduces the violation decomposition for targeted diagnosis of incompatibility. Section 3-6 formulates the cost function and Section 3-7 describes the optimization procedure. Section 3-8 outlines network-level inference with optimized priors. Finally, Section 3-9 outlines the validation strategy.

3-1 Problem setting and assumptions

This section describes the problem setting and modeling assumptions that form the basis of the proposed approach. The design focuses on two aspects: the proposed prior design setting that supports the approach, and the modeling assumptions that govern its formalization.

3-1-1 Prior design setting

In many practical applications of BBNs, the graphical structure is known or assumed fixed, while the available target data for CPT parameter learning is limited. In such settings, posterior inference becomes highly sensitive to the choice of prior distribution. When little data is available, the prior can dominate the posterior; if the prior is poorly chosen, the resulting estimates may become unstable, overconfident, or inconsistent with known domain relationships.

A standard baseline in this setting is to assume a uniform prior over the admissible parameter space [68, 38]. When target-side uncertainty and domain-informed constraints are available, this idea can be refined by considering a uniform distribution restricted to the corresponding feasible parameter region. Such a prior has the advantage of remaining fully consistent with available information. However, it treats all admissible parameter configurations as equally likely and therefore fails to encode any preference within the feasible set. In particular, when the feasible region exhibits geometric structure induced by cross-row constraints, a uniform prior over that region cannot distinguish between central, representative configurations and fragile solutions near the constraint boundaries. Moreover, experts generally cannot specify full prior hyperparameter vectors $\alpha_{ij1}, ..., \alpha_{ijr_i}$, and are typically more comfortable providing qualitative knowledge [16, 41, 40] such as

$$\mathbb{P}(I = \mathrm{T} \mid H_1 = \mathrm{F}, H_2 = \mathrm{T}) > \mathbb{P}(I = \mathrm{T} \mid H_1 = \mathrm{F}, H_2 = \mathrm{F}).$$

Such statements impose constraints on the parameter space but do not naturally correspond to a tractable parametric prior. This motivates the approach proposed in this thesis: rather than assuming uniformity within the feasible region, a parametric prior is optimized to concentrate probability mass on representative parameter configurations while remaining consistent with both the data-driven uncertainty and the imposed cross-row constraints. The primary setting considered is one in which no source domain data is available and the prior is constructed entirely from target-side information. When source domain data is available, it can alternatively serve as the basis for a candidate prior selection, which is evaluated comparatively in Section 3-4. The formalization of the primary setting begins in Section 3-2.

3-1-2 Modeling assumptions

The following assumptions govern the proposed framework.

Assumption 1. *The graphical structure of the BBN is known and fixed. Only the CPT parameters are subject to uncertainty and estimation.*

Assumption 2. *All CPT nodes are binary, taking values in $\{0, 1\}$. Each CPT row j defines a Bernoulli probability mass function with entries θ^0_{ij} and $\theta^1_{ij} = 1 - \theta^0_{ij}$, leaving a single free parameter $\theta^1_{ij} = \mathbb{P}(X_i = 1 \mid \Pi_i = j) \in (0, 1)$ per row. The framework is developed for this binary case; extension to nodes with more than two states via a Dirichlet-Multinomial formulation is left for future work.*

Assumption 3. *Each CPT row parameter θ_{ij} is modeled using a Beta distribution. This choice is motivated by conjugacy with the Binomial likelihood, which permits closed-form posterior updating [26, 37].*

Assumption 4. *Parameters corresponding to different CPT rows are treated as locally independent during posterior updating. This is consistent with the standard factorized structure of BBN parameter learning, in which each conditional distribution is estimated separately given its parent configuration [26].*

Assumption 5. *Parent variable states are observable in the target dataset. When a parent configuration has no observations in the calibration split ($N_{ij} = 0$), the Wilson interval is set to $[0, 1]$, consistent with the limiting behavior established in Section 2-2-4. This imposes no data-driven constraint on that row; the feasible region is then shaped by the cross-row constraints alone, and the posterior equals the prior since no likelihood contribution exists for that row.*

Assumption 6. *The imposed domain constraints are satisfied by the true underlying CPT parameters. This ensures that the feasible region contains the true parameter values and that the proposed framework does not systematically bias the prior away from the truth.*

Assumption 7. *Constraints are only incorporated during prior construction. Posterior inference is performed using standard conjugate updates without enforcing cross-row constraints at inference time. This separation ensures tractability of the posterior and is consistent with the view that constraints inform the prior whereas the likelihood purely reflects the observed data.*

Assumption 8. *The observed target data are assumed to be generated from a fixed underlying CPT.*

Together, these assumptions define the scope within which the proposed framework operates. Under these conditions, the feasible parameter region is constructed in Section 3-2, metrics for candidate prior evaluation are introduced in Section 3-3, and the optimization procedure is formalized in Section 3-6 and 3-7.

3-2 Feasible parameter region from target information

This section describes the construction of the feasible parameter region $\mathcal{F}_i$, which serves as the foundation for the prior evaluation and optimization frameworks. The region is defined by two distinct sources of information: row-wise uncertainty intervals derived from limited target data, and qualitative cross-row constraints encoding domain knowledge. Their intersection yields a structured subset of the parameter space that contains all parameter configurations consistent with both sources.

3-2-1 Representation of target uncertainty

Let $\mathcal{D}_{\text{target}} = \{x^{(1)}, \ldots, x^{(N_{\text{target}})}\}$ denote a dataset sampled from an unknown target distribution P_{target}. We assume a fixed BBN structure and focus on estimating the CPT parameters.

Consider a node X_i with discrete state space $\mathcal{X}_i$ and parent set $\text{Pa}(X_i)$. Let $j \in \mathcal{I}_i$ index the parent configurations, and let $r_i = |\mathcal{X}_i|$ denote the number of states of X_i. The CPT parameters are given by

$$\theta_{ijk} = \mathbb{P}(X_i = k \mid \text{Pa}(X_i) = j), \quad k = 1, \ldots, r_i, \tag{3-1}$$

with the simplex constraint $\sum_{k=1}^{r_i} \theta_{ijk} = 1$ for each j. Given the target dataset, let N_{ij} denote the number of observations corresponding to parent configuration j, and let N_{ijk} denote the number of observations in which $X_i = k$ under configuration j.

Due to limited sample sizes, point estimates of the CPT parameters θ_{ijk} can be unreliable. The proposed approach therefore represents target-side uncertainty through interval estimates rather than single point estimates. For each parameter, an interval is constructed as

$$\theta_{ijk} \in [L_{ijk},\, U_{ijk}], \tag{3-2}$$

where L_{ijk} and U_{ijk} are derived from the observed counts (N_{ijk}, N_{ij}).

In the binary case ($r_i = 2$), each CPT row contains one free parameter θ_{ij}, with interval $\theta_{ij} \in [L_{ij}, U_{ij}]$. The interval bounds are constructed using Wilson score intervals, which provide improved coverage properties compared to the standard normal approximation, especially for small sample sizes and probabilities near the boundaries [43, 44]. For a Bernoulli parameter with empirical proportion $\hat{p}_{ij} = \frac{N^1_{ij}}{N_{ij}}$, the Wilson score interval is $[L_{ij}, U_{ij}]$, where

$$L_{ij} = \frac{\hat{p}_{ij} + \dfrac{z^2}{2N_{ij}} - z\sqrt{\dfrac{\hat{p}_{ij}(1-\hat{p}_{ij})}{N_{ij}} + \dfrac{z^2}{4N_{ij}^2}}}{1 + \dfrac{z^2}{N_{ij}}}, \tag{3-3}$$

$$U_{ij} = \frac{\hat{p}_{ij} + \dfrac{z^2}{2N_{ij}} + z\sqrt{\dfrac{\hat{p}_{ij}(1-\hat{p}_{ij})}{N_{ij}} + \dfrac{z^2}{4N_{ij}^2}}}{1 + \dfrac{z^2}{N_{ij}}}, \tag{3-4}$$

and z denotes the standard normal quantile of the desired confidence level.

The width $\Delta_{ij} = U_{ij} - L_{ij}$ of each Wilson interval decreases monotonically in N_{ij} for fixed α and $\hat{p}_{ij}$ [69], so the row-wise uncertainty intervals contract as target observations accumulate, leading to a progressively smaller credal box $\mathcal{B}_i$ after aggregation. In the data-scarce regime targeted by this thesis, wide intervals reflect parameter uncertainty rather than model misspecification.

Under Assumption 4, each row-wise interval is estimated separately from its own sufficient statistics. The joint uncertainty set for node X_i is therefore represented as the Cartesian product:

$$\mathcal{B}_i = \prod_{j \in \mathcal{I}_i} [L_{ij},\, U_{ij}]. \tag{3-5}$$

In the binary case ($r_i = 2$), each CPT row contains one free parameter $\theta_{ij} \in [0, 1]$, so $\mathcal{B}_i$ becomes a hyperrectangle in $[0, 1]^{|\mathcal{I}_i|}$. Each interval represents the admissible range of a CPT row parameter supported by the target data, and their Cartesian product defines the set of all parameter vectors consistent with these row-wise uncertainty bounds. Following the terminology of imprecise probability theory, such hyper-rectangular uncertainty sets are referred to as credal boxes in this research [18].

Across the full network, the joint uncertainty set is

$$\mathcal{B} = \prod_i \mathcal{B}_i \subseteq \Theta, \tag{3-6}$$

where Θ denotes the full parameter space of the BBN. This set captures epistemic uncertainty induced by limited target data, prior to incorporating any additional domain-informed constraints.

3-2-2 Qualitative cross-row constraints

In addition to data-driven uncertainty, domain knowledge can often be expressed as qualitative relationships between CPT parameters across different parent configurations. Such constraints impose additional structure on the CPT and restrict the admissible parameter space beyond what the interval estimates alone capture.

Definition 1. *Let θ_{ijk} and $\theta_{i'j'k}$ be parameters corresponding to two (possibly distinct) parent configurations (i, j) and (i', j') of a CPT. A constraint involving both parameters is called a* cross-row constraint *if $j \neq j'$, i.e., if it relates parameters from different rows of the CPT. Such constraints introduce dependencies between parameters that would otherwise be treated as independent under standard BBN parameter-learning assumptions.*

Cross-row constraints take one of the following forms in this research:

- **Monotonicity constraints:** enforcing that a parameter increases or decreases with respect to a parent variable, e.g. $\theta_{ij_1} \leq \theta_{ij_2}$ when configuration j_2 represents a "higher" state than j_1.
- **Gap constraints:** enforcing a minimum separation between parameters, e.g. $\theta_{ij_1} - \theta_{ij_2} \geq \delta$ for some $\delta > 0$.
- **Closeness constraints:** bounding the maximum difference between parameters, e.g. $\theta_{ij_1} - \theta_{ij_2} \leq \varepsilon$ for some $\varepsilon > 0$, preventing any single parent configuration from dominating another by more than a specified tolerance.

More generally, these constraints can be expressed as a finite set of inequality functions:

$$g_\ell(\theta) \leq 0, \quad \ell = 1, \ldots, K, \tag{3-7}$$

where each g_ℓ encodes a cross-row relation between parameters. Let K denote the number of active constraints under the chosen constraint mode. The constraints are independent of the observed data and reflect prior qualitative knowledge about the modeled domain. They play a key role in shaping the geometry of the feasible region.

Unlike interval bounds, which act independently on individual CPT rows, cross-row constraints introduce dependencies between parameters across different parent configurations. Consequently, feasibility can no longer be assessed on a row-by-row basis. A candidate parameter vector may satisfy all row-wise interval constraints while still violating the joint qualitative structure imposed by the cross-row constraints. The admissible parameter space is therefore determined by the joint interaction of interval bounds and cross-row constraints rather than by individual rows in isolation.

3-2-3 Joint feasible-region construction

The feasible parameter region is obtained by intersecting the data-driven credal box with the set of parameter vectors satisfying the domain-informed cross-row constraints. Let $\Theta \subseteq [0, 1]^n$ denote the parameter space of interest. The joint feasible region of a node X_i is defined as

$$\mathcal{F}_i = \{\, \theta \in \mathcal{B}_i \mid g_\ell(\theta) \leq 0 \;\; \forall\, \ell = 1, \ldots, K \,\}. \tag{3-8}$$

This region contains all parameter configurations that are simultaneously consistent with the observed target data and the imposed domain-informed constraints, and serves as the feasible set for constructing and evaluating candidate prior distributions in the following sections.

The formulation of a feasible parameter set through inequality constraints on CPT parameters follows the constrained parameter-learning framework of [49], where the space of admissible parameters is defined as the intersection of equality and inequality constraint sets derived from domain knowledge.

The cross-row constraints introduce dependencies between parameters, so $\mathcal{F}_i$ is in general no longer a box but a structured subset of $\mathcal{B}_i$. Since $\mathcal{B}_i$ is itself a bounded convex set, and each linear cross-row constraint defines a closed halfspace, $\mathcal{F}_i$ is the bounded intersection of finitely many halfspaces. By standard results in convex geometry, such a set is a bounded convex polytope [70].

Throughout this thesis, the data-scarce regimes ensure that $\mathcal{F}_i$ is non-empty in practice: the credal box $\mathcal{B}_i$ is wide relative to the constraint set, so the intersection remains non-empty. This is treated as a regularity condition of the framework.

Although the feasible region can in principle be defined over the full network, the proposed approach operates locally at the level of individual nodes or groups of parameters sharing cross-row constraints. This is consistent with the factorized structure of BBNs and ensures computational tractability [26].

3-3 Metrics for prior-region compatibility

The feasible parameter region $\mathcal{F}_i$ constructed in Section 3-2 provides a structured subset of the parameter space that is consistent with both target-side uncertainty and domain-informed constraints. To evaluate how well a candidate prior distribution aligns with this region, two complementary metrics are introduced: the prior mass assigned to the feasible region, and the expected degree of constraint violation.

The underlying diagnostic perspective is motivated by the principle that a prior distribution should not concentrate probability on parameter values that are inconsistent with the available evidence. Prior-data conflict arises when the prior assigns most of its mass to parameter values for which the observed data is surprising, and detecting such conflict is an essential step before drawing inferences from a Bayesian model [71]. The present approach adapts this diagnostic perspective to a constrained parameter space setting. Rather than recovering the region of data-consistent parameter values implicitly from the likelihood, the feasible region $\mathcal{F}_i$ makes this set explicit: a parameter vector is inadmissible not only if it conflicts with the observed target counts, but also if it violates qualitative structural constraints encoding domain knowledge. This broader notion of feasibility justifies two complementary metrics, since a single mass-based diagnostic cannot decompose incompatibility into its data-driven and constraint-driven sources.

Together, these metrics characterize prior-region compatibility from separate perspectives. The mass metric captures how much probability the prior places on admissible parameter values. The violation metric quantifies how severely the prior violates feasibility constraints, distinguishing between the frequency and magnitude of violations.

These metrics serve a dual purpose in the proposed framework. In the prior evaluation setting, $\mathbb{P}_s$ denotes a fixed candidate source prior, and the metrics quantify its compatibility with $\mathcal{F}_i$. In the prior construction setting, $\mathbb{P}_\phi$ denotes a parametric prior distribution with hyperparameters ϕ to be optimized; the same metrics then define terms in the cost function introduced in Section 3-6. This unified perspective allows the same metrics to support both the evaluation of existing source priors and the construction of new priors from target-side information.

Since the feasible region $\mathcal{F}_i$ does not admit a closed-form analytical description in general, both metrics are estimated via Monte Carlo sampling. This sample average approximation is a

standard technique for stochastic optimization problems in which expectations over constrained sets cannot be evaluated analytically [72].

3-3-1 Prior mass in the feasible region

The first metric measures the total probability that a candidate prior $\mathbb{P}$ assigns to the feasible region. Formally, it is defined as the integral of the prior density over $\mathcal{F}_i$:

$$m(\mathbb{P}) = \int_{\mathcal{F}_i} \mathbb{P}(\theta)\, d\theta = \mathbb{P}(\theta \in \mathcal{F}_i). \tag{3-9}$$

A prior with high mass in $\mathcal{F}_i$ assigns substantial probability to parameter configurations that are simultaneously consistent with the observed target data and the imposed domain constraints [71]. Conversely, a prior with low mass concentrates probability outside the feasible region, producing parameter configurations that are likely to violate either the data-driven uncertainty bounds or the structural constraints.

Since the integral of a Beta density over the polytope $\mathcal{F}_i$ does not admit a closed-form expression in general, the mass is written as the expectation of a feasibility indicator,

$$m(\mathbb{P}) = \mathbb{E}_{\theta\sim\mathbb{P}}\big[\mathbf{1}[\theta \in \mathcal{F}_i]\big], \tag{3-10}$$

and estimated via Monte Carlo sampling. Given S independent draws $\theta^{(1)}, \ldots, \theta^{(S)} \sim \mathbb{P}$, the estimator is

$$\hat{m}(\mathbb{P}) = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\big[\theta^{(s)} \in \mathcal{F}_i\big], \tag{3-11}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. This estimator is unbiased and converges to $m(\mathbb{P})$ as $S \to \infty$ by the law of large numbers. Feasibility of each sample is assessed by checking both the row-wise box constraints defining $\mathcal{B}_i$ and the cross-row constraints $g_\ell(\theta) \le 0$.

3-3-2 Constraint violation metrics

Prior mass alone does not explain *why* a distribution is incompatible with the feasible region, nor does it quantify *how severely* infeasible samples violate the constraints. The idea of quantifying the degree to which CPT parameters violate qualitative constraints through a penalty function has been used in constrained BBN parameter learning [60, 52]. To address this, constraint violations are explicitly quantified using hinge-type penalty functions of the form $\max(g_\ell(\theta), 0)$, which are zero when the constraint is satisfied and grow linearly with the degree of violation otherwise.

For each parameter vector θ, the total violation is defined as

$$V(\theta) = \sum_{j} V_{\text{box},j}(\theta) + \sum_{\ell=1}^{K} V_{c,\ell}(\theta), \tag{3-12}$$

where the row-wise box constraint violations are given by

$$V_{\text{box},j}(\theta) = \max(L_{ij} - \theta_{ij},\, 0) + \max(\theta_{ij} - U_{ij},\, 0), \tag{3-13}$$

and the cross-row constraint violations are penalized as

$$V_{c,\ell}(\theta) = \max\big(g_\ell(\theta),\, 0\big), \tag{3-14}$$

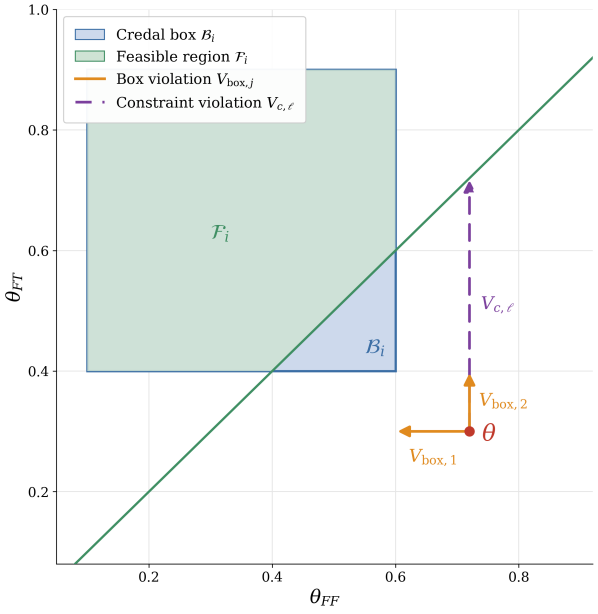


Figure 3-1: Geometric illustration of the two violation components in Equation (3-12) for a two-row CPT. The credal box $\mathcal{B}_i$ (blue) is defined by the row-wise Wilson intervals. The feasible region $\mathcal{F}_i$ (green) is the intersection of $\mathcal{B}_i$ with the monotonicity constraint $\theta_{FT} \geq \theta_{FF}$. A parameter vector θ outside $\mathcal{F}_i$ incurs box violations $V_{\text{box},j}$ measuring how far θ extends beyond each Wilson interval bound (orange), and a cross-row constraint violation $V_{c,\ell} = \max(\theta_{FF} - \theta_{FT}, 0)$ equal to the algebraic constraint surplus (purple).

with $g_\ell(\theta) \leq 0$.

Figure 3-1 illustrates the two components geometrically for a two-row CPT under one monotonicity constraint.

Each term contributes a non-negative penalty proportional to the magnitude of the violation: $V(\theta) = 0$ if and only if $\theta \in \mathcal{F}_i$, and $V(\theta) > 0$ otherwise. These hinge penalties can be interpreted as computationally efficient proxies for the distance-to-set measure

$$\text{dist}(\theta, \mathcal{F}_i) = \inf_{\theta' \in \mathcal{F}_i} \|\theta - \theta'\|, \tag{3-15}$$

which provides the tightest continuous measure of infeasibility but is costly to evaluate in general [70].

To evaluate a prior distribution $\mathbb{P}$ under this metric, the expected violation is computed:

$$\mathbb{E}_{\theta \sim \mathbb{P}}\big[V(\theta)\big], \tag{3-16}$$

which measures the average degree of infeasibility across all parameter configurations drawn from the prior. A prior with low expected violation concentrates probability on configurations that are either feasible or close to the feasible boundary.

The hinge form is preferred over a binary feasibility indicator $\mathbf{1}[\theta \notin \mathcal{F}_i]$ for two reasons. First, a binary indicator cannot distinguish a mild violation from a severe one, whereas the hinge produces a continuous quantity proportional to the degree of infeasibility, making the expected violation a meaningful compatibility metric rather than merely a feasibility rate. Second, since $V(\theta)$ decomposes additively into per-row box violations and per-constraint cross-row violations (Equation 3-12), linearity of expectation means the expected violation inherits this decomposition exactly, so each row and each constraint contributes an independently interpretable share to the global incompatibility score. This additive structure is exploited in Section 3-5-1.

For diagnostic purposes, the conditional expectation

$$\mathbb{E}_{\theta\sim\mathbb{P}}[V(\theta) \mid \theta \notin \mathcal{F}_i] \tag{3-17}$$

can also be computed by restricting the average to infeasible samples. This quantity captures the average severity of violations given that infeasibility occurs, and is useful for identifying whether a prior is marginally or severely inconsistent with $\mathcal{F}_i$. It does not enter the cost function but serves as a secondary diagnostic when evaluating candidate source priors.

3-4 Source domain prior evaluation

The compatibility metrics introduced in Section 3-3 serve a dual purpose in the proposed framework. In the prior evaluation setting considered here, the same metrics are applied to assess whether a candidate source prior, derived from an external domain such as a related geographical region, a historical dataset, or a comparable operational context, is sufficiently consistent with the target-side feasible region to serve as an informative prior for downstream inference. Through this, it can be evaluated whether the metrics are able to distinguish well-aligned from poorly-aligned source priors, and how their discriminative power varies with target sample size and constraint tightness.

This evaluation setting arises naturally in transfer learning for BBNs, where source domain knowledge is available but its relevance to the target domain is uncertain. Rather than accepting or rejecting a source prior wholesale, the proposed framework quantifies the degree of compatibility, enabling systematic selection and, where relevant, identification of the specific parameters driving incompatibility.

3-4-1 Source prior representation

It is assumed that prior knowledge is available in the form of one or more candidate source distributions over the CPT parameters:

$$\mathbb{P}_s, \quad s \in \mathcal{S}, \tag{3-18}$$

where $\mathcal{S}$ indexes different sources. In this thesis, each source prior $\mathbb{P}_s$ is represented as a row-wise product of Beta distributions over the binary CPT row parameters, consistent with Assumption 3. Each source prior $\mathbb{P}_s$ encodes beliefs about the CPT parameters, but may be inconsistent with the target domain, partially informative, or systematically biased due to domain shift.

3-4-2 Compatibility-based source prior ranking

For each candidate source prior $\mathbb{P}_s$, two compatibility scores are computed against the target-side feasible region $\mathcal{F}_i$:

- $\hat{m}(\mathbb{P}_s)$: prior mass in the feasible region, estimated via Equation (3-11).
- $\mathbb{E}_{\theta\sim\mathbb{P}_s}[V(\theta)]$,: expected constraint violation, estimated via Equation (3-16).

These two scores characterize compatibility from distinct perspectives: mass captures the overall proportion of prior probability assigned to admissible parameter configurations, while violation captures the severity of infeasibility for configurations that fall outside $\mathcal{F}_i$. A source prior is

considered well-aligned with the target domain if it achieves high prior mass and low expected violation. A source prior that performs poorly on both metrics is more likely to induce negative transfer and should be excluded from downstream inference. A source prior that shows moderate scores warrants further inspection via the violation decomposition described in Section 3-5, which identifies the specific rows and constraints driving the incompatibility.

3-4-3 Limitations of compatibility-based evaluation

The compatibility metrics evaluate a source prior against the target-side feasible region, which is itself constructed from limited target data. Under severe data scarcity, the feasible region $\mathcal{F}_i$ is wide, and multiple source priors may appear equally compatible even when their true alignment with the target domain differs substantially. In this regime, the metrics correctly reflect the limited information content of the available data rather than failing as diagnostic tools: the feasible region defines a set of acceptable priors rather than a unique optimum, and the inability to discriminate between near-equivalent sources is an inherent consequence of data scarcity.

Discriminative power increases as target data accumulates and as cross-row constraints tighten the feasible region. Domain-informed structural constraints therefore play a dual role: they encode qualitative domain knowledge and simultaneously amplify the informativeness of the compatibility metrics. The experimental validation of these properties is reported in Section 4-2.

3-5 Diagnostic violation decomposition

The compatibility metrics of Section 3-3 produce scalar summaries of how well a candidate prior aligns with the feasible region $\mathcal{F}_i$ as a whole. While these global scores are sufficient for ranking and selection, they do not explain *where* incompatibility arises or *which* aspects of the prior are responsible for it. This section introduces a decomposition of the global violation score into row-level and constraint-level contributions, enabling targeted diagnosis of incompatibility.

3-5-1 Decomposition of global incompatibility

The total violation $V(\theta)$ decomposes naturally into two additive components:

$$V(\theta) = \underbrace{\sum_j V_{\text{box},j}(\theta)}_{\text{row-wise violation}} + \underbrace{\sum_\ell V_{c,\ell}(\theta)}_{\text{cross-row violation}} . \tag{3-19}$$

Taking the expectation over a candidate prior $\mathbb{P}_s$ yields a decomposed expected violation:

$$\mathbb{E}_{\theta\sim\mathbb{P}_s}[V(\theta)] = \sum_j \mathbb{E}[V_{\text{box},j}(\theta)] + \sum_\ell \mathbb{E}[V_{c,\ell}(\theta)], \tag{3-20}$$

where:

- $\mathbb{E}[V_{\text{box},j}(\theta)]$ measures the expected violation of the data-driven Wilson interval for row j. A large value indicates that the source prior places substantial mass outside the uncertainty interval supported by the target data for that row, suggesting a row-level location mismatch between the source and target domains.

- $\mathbb{E}[V_{c,\ell}(\theta)]$ measures the expected violation of cross-row constraint g_ℓ. A large value indicates that the source prior frequently samples parameter configurations that violate the structural ordering encoded by that constraint, independently of whether individual rows fall within their Wilson intervals.

A source prior may pass row-wise assessment, with each individual row within its Wilson interval, while still violating a cross-row monotonicity constraint. This is precisely the limitation of column-wise compatibility screening in MPL-TC, discussed in Section 2-4. The decomposition detects both types of incompatibility and attributes global violation to its specific sources.

3-5-2 Identifying critical CPT parameters

The row-level and constraint-level violation profiles identify which CPT rows and structural relationships contribute most to incompatibility. Formally, the row-wise violation profile is defined as the vector

$$\boldsymbol{v}_{\text{box}} = \big(\mathbb{E}[V_{\text{box},1}(\theta)], \ldots, \mathbb{E}[V_{\text{box},J}(\theta)]\big), \tag{3-21}$$

and the constraint-level violation profile is

$$\boldsymbol{v}_c = \big(\mathbb{E}[V_{c,1}(\theta)], \ldots, \mathbb{E}[V_{c,K}(\theta)]\big). \tag{3-22}$$

Both vectors are estimated via Monte Carlo sampling from $\mathbb{P}_s$ and provide a fine-grained diagnostic of where incompatibility is concentrated. As illustrated in Section 4-2-5, violation is typically concentrated on a small number of rows and constraints when domain shift is localized, reflecting the structure of the specific mismatch rather than a universal property of prior incompatibility.

This localization has a direct operational interpretation. An analyst reviewing a candidate source prior can focus refinement efforts on the specific rows and constraints flagged by the decomposition, rather than discarding the source prior entirely. For example, if the violation profile indicates that a single row j^* accounts for the majority of box violation while the remaining rows are well-aligned, the analyst can update the source prior for that row j^* alone (perhaps by eliciting a revised estimate for that parent configuration) without discarding the remainder of the source prior.

In the source-unavailable setting of Section 3-6, explicit decomposition is unnecessary because the optimizer constructs a prior directly within the feasible-region framework. When a candidate source prior exists and must be assessed before use, however, the decomposition provides a structured basis for targeted expert review: an analyst can identify precisely which rows and constraints drive incompatibility rather than relying solely on a single aggregate score. The additional experimental validation of these localization properties is reported in Section 4-2.

Up to this point, the compatibility metrics have been introduced as diagnostic tools for evaluating existing priors. The remainder of this chapter considers a second use case. Rather than evaluating an externally specified prior, the same metrics are incorporated directly into a cost function to construct a prior from target-side information alone. This shifts the role of the metrics from prior evaluation (RQ2) to prior construction (RQ3A).

3-6 Bayesian prior optimization

The metrics introduced in Section 3-3 provide a basis for evaluating how well a candidate prior aligns with the feasible parameter region $\mathcal{F}_i$. This section translates those metrics into a cost

function, formulates the parametric prior family over which optimization is performed, and derives the combined cost function used for prior construction.

The central idea is to construct a prior distribution whose probability mass is concentrated on parameter configurations that are consistent with both the data-driven uncertainty and the domain-informed constraints, while avoiding overconfidence relative to the information available from the target data. The feasible region itself does not uniquely determine a prior distribution. In general, many distinct prior distributions can satisfy the same constraints and assign substantial probability mass to the same feasible region. The optimization problem therefore seeks a prior that is not only feasible, but also representative and appropriately calibrated with respect to the uncertainty supported by the target data.

The compatibility metrics introduced in Section 3-3 naturally lend themselves to a penalty-based optimization formulation. Rather than enforcing hard feasibility constraints, violations are incorporated as penalty terms in a smooth unconstrained cost function. This choice is motivated by the fact that constraint satisfaction is assessed via Monte Carlo sampling: feasibility can only be evaluated on a finite set of draws, making strict constraint enforcement brittle and sensitive to sampling fluctuations [73]. A penalty-based formulation instead allows the optimizer to make gradual progress toward feasible regions and to trade off feasibility, concentration, and representativeness in a systematic way. The resulting nonconvex stochastic cost function is minimized using derivative-free optimization methods described in Section 3-7.

3-6-1 Parametric prior family

Under Assumption 3, each binary CPT row parameter θ_j is modeled using a Beta prior [33, 26, 34],

$$\theta_j \sim \mathrm{Beta}(\alpha_j, \beta_j). \tag{3-23}$$

Rather than optimizing directly over the shape parameters (α_j, β_j), the prior is reparameterized in terms of a mean μ_j and concentration τ_j:

$$\alpha_j = \mu_j \tau_j, \qquad \beta_j = (1-\mu_j)\tau_j, \qquad \mu_j \in (0,1),\ \tau_j > 0. \tag{3-24}$$

This separates the two interpretable roles of the prior:

- μ_j controls the *location* of the prior, i.e., the expected value $\mathbb{E}[\theta_j] = \mu_j$,
- τ_j controls the *concentration*, with prior variance

$$\mathrm{Var}(\theta_j) = \frac{\mu_j(1-\mu_j)}{\tau_j + 1}, \tag{3-25}$$

  so that larger τ_j yields a more concentrated prior with smaller variance [74].

The optimization variables are therefore $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_J)$, and the joint prior over the full CPT parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ factorizes as

$$p(\boldsymbol{\theta}) = \prod_{j=1}^{J} \mathrm{Beta}(\theta_j \mid \mu_j\tau_j,\ (1-\mu_j)\tau_j). \tag{3-26}$$

The mean concentration parameterization provides a more interpretable and structurally meaningful optimization space than the shape parameterization. The optimization over μ_j determines where the prior places its mass relative to the feasible region: if μ_j lies inside $\mathcal{F}_i$, the prior is well-aligned with the data and constraints; if μ_j lies outside, the prior induces systematic constraint

violations. The optimization over τ_j determines how strongly this belief influences the posterior: large τ_j increases the impact of prior misspecification, while small τ_j allows the likelihood to dominate. This separation allows the framework to distinguish between *location mismatch* and *overconfidence* as distinct failure modes of a candidate prior and can therefore directly control how strongly prior influences the posterior.

Although the prior in Equation (3-26) factorizes across rows, the optimization is not performed independently per row. During optimization, samples $\boldsymbol{\theta}^{(s)} = (\theta_1^{(s)}, \dots, \theta_J^{(s)})$ are drawn independently per row from the product prior, assembling joint CPT candidate vector. Each sampled vector represents an entire candidate CPT, and the cost function evaluates its compatibility with the joint feasible region $\mathcal{F}_i \subseteq [0,1]^J$, which couples all rows through the cross-row constraints. This prevents independent row-wise collapse and encourages globally coherent CPT configurations.

3-6-2 Distance to representative target configuration

The prior mass and violation metrics assess compatibility with the feasible region in terms of feasibility alone. However, $\mathcal{F}_i$ has geometric structure: it contains configurations ranging from central and representative ones to those near the boundary that are technically feasible but fragile. Two priors may assign identical mass to $\mathcal{F}_i$ while differing substantially in where their probability is concentrated within it. A prior concentrated near the boundary of $\mathcal{F}_i$ is fragile with respect to perturbations of the feasible region itself. When additional target observations arrive and Wilson intervals narrow, or when an additional domain constraint is imposed, boundary solutions may become infeasible while interior solutions remain admissible. A prior centered in the interior of $\mathcal{F}_i$ is therefore more robust to such changes in the available evidence and constraint set.

A third metric measures proximity to the interior of $\mathcal{F}_i$ rather than mere feasibility. The reference point is the projected center of $\mathcal{F}_i$, defined row-wise as the midpoint of the feasible range of each CPT row:

$$\theta_{c,j} = \tfrac{1}{2}(a_j + b_j), \tag{3-27}$$

where a_j and b_j are the minimum and maximum values that θ_j can attain within $\mathcal{F}_i$:

$$a_j = \min_{\theta \in \mathcal{F}_i} \theta_j, \qquad b_j = \max_{\theta \in \mathcal{F}_i} \theta_j.$$

Since $\mathcal{F}_i$ is defined by linear constraints over a bounded box, each of these is a linear program and is solved exactly. The resulting center $\theta_c = (\theta_{c,1}, \dots, \theta_{c,J})$ represents the midpoint of the range that each CPT row can take while remaining jointly consistent with both the data-driven uncertainty bounds and the structural constraints. It therefore reflects the interior of $\mathcal{F}_i$ rather than the interior of the wider credal box $\mathcal{B}_i$, and correctly accounts for the tightening of individual row ranges induced by cross-row constraints.

The distance metric is then defined as the expected L_1 deviation of sampled parameter vectors from this center:

$$D(\mathbb{P}) = \mathbb{E}_{\theta \sim \mathbb{P}}\big[\|\theta - \theta_c\|_1\big] = \mathbb{E}_{\theta \sim \mathbb{P}}\left[\sum_{j=1}^{J} \|\theta_j - \theta_{c,j}\|\right]. \tag{3-28}$$

A prior with low $D(\mathbb{P})$ concentrates mass near the interior of $\mathcal{F}_i$, favoring stable and representative parameter configurations. A prior with high $D(\mathbb{P})$ is displaced from this center, either

because its mean is misaligned with the feasible region or because its spread is large relative to the region's extent. The L_1 norm is adopted for its interpretability and robustness: it does not disproportionately penalize individual rows with large deviations and scales linearly with the number of CPT rows.

As with the other metrics, $D(\mathbb{P})$ is estimated via Monte Carlo sampling:

$$\hat{D}(\mathbb{P}) = \frac{1}{S}\sum_{s=1}^{S} \|\theta^{(s)} - \theta_c\|_1, \qquad \theta^{(s)} \sim \mathbb{P}.$$

The center θ_c is recomputed from the current calibration data and constraint mode at each experimental run, so it adapts to the specific feasible region encountered. This metric enters as the centering term in the cost function formulated in Section 3-6, where it discourages priors displaced from the representative interior of $\mathcal{F}_i$.

3-6-3 Base cost function

The base cost function promotes compatibility of the prior with the feasible region $\mathcal{F}_i$ by combining the two metrics introduced in Section 3-3, augmented with a centering distance regularization term. It balances three complementary goals: penalizing constraint violations, maximizing prior mass inside $\mathcal{F}_i$, and encouraging proximity to the projected center of feasible region.

Beyond ranking candidate priors, a continuous violation measure also provides a more informative optimization signal than a binary feasibility indicator, which would contribute zero gradient almost everywhere and give the optimizer no guidance toward the feasible-region interior.

Formally, the base cost function is defined as

$$\mathcal{L}_{\text{base}}(\boldsymbol{\mu}, \boldsymbol{\tau}) = w_v\, \mathbb{E}_{\theta\sim\mathbb{P}_\phi}\big[V(\theta)\big] + w_d\, \mathbb{E}_{\theta\sim\mathbb{P}_\phi}\big[\|\theta - \theta_c\|_1\big] - w_m\, \mathbb{E}_{\theta\sim\mathbb{P}_\phi}\big[\mathbf{1}[\theta \in \mathcal{F}_i]\big], \tag{3-29}$$

where $\mathbb{P}_\phi$ denotes the parametric prior $p(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\tau})$, $V(\theta)$ is the total constraint violation defined in Equation (3-12), θ_c is the feasible-region center defined in Equation (3-27), and $w_v, w_d, w_m > 0$ are weights controlling the relative importance of each term.

The centering term $D(\mathbb{P})$, introduced in Section 3-6-2, provides a continuous penalty signal that favors interior configurations of $\mathcal{F}_i$ over boundary-adjacent ones, complementing the violation and mass terms which provide no geometric signal within the feasible region.

The violation term penalizes infeasible samples proportionally to their degree of infeasibility, guiding the prior mean toward the interior of $\mathcal{F}_i$. The mass term rewards probability assigned to feasible configurations, reinforcing the violation term from a different angle. The centering term discourages solutions displaced toward the boundary of the feasible region, favoring stable and representative configurations over fragile boundary solutions. Note that the binary indicator remains the natural estimator for the mass term, where the quantity of interest is precisely the probability of feasibility; the continuous hinge is used specifically for the violation term, which must supply an informative optimization signal toward the feasible region rather than a feasibility rate.

Although the cost function promotes feasibility and representativeness, it does not explicitly control the concentration of the prior. In the absence of additional regularization, the optimizer tends to drive τ_j toward very large values: since increasing τ_j concentrates the prior mass on a narrow region, it can artificially improve all three terms in $\mathcal{L}_{\text{base}}$ without reflecting any genuine increase in available information. This pathological behavior motivates the target-uncertainty regularization introduced in the following subsection.

3-6-4 Target-uncertainty regularization of prior concentration

The regularization term controls whether the concentration of the optimized prior is commensurate with the uncertainty supported by the target data. Its goal is not to penalize large values of τ_j directly, but to prevent the prior from becoming more concentrated than is warranted by the available evidence.

Target-side variance proxy

The target uncertainty for row j is represented through the Wilson interval $[L_j, U_j]$, whose width reflects the epistemic uncertainty remaining after observing a limited number of target samples. To compare prior uncertainty and target uncertainty on a common scale, a heuristic measure of uncertainty implied by the Wilson interval is converted into a variance proxy.

For a symmetric confidence interval of the form $\hat{p}_j \pm z\sigma_j$, the variance can be recovered from the interval half-width by $\sigma_j^2 = (\omega_j/z)^2$. Wilson intervals are generally asymmetric under sparse data [45], yielding different lower and upper half-widths relative to the empirical proportion $\hat{p}_j = n_j/N_j$. Applying the relationship above separately to the lower and upper Wilson half-widths yields the variance estimates

$$(\sigma_j^-)^2 = \left(\frac{\hat{p}_j - L_j}{z}\right)^2, \qquad (\sigma_j^+)^2 = \left(\frac{U_j - \hat{p}_j}{z}\right)^2. \tag{3-30}$$

In the absence of additional information regarding the relative importance of the lower and upper interval widths, equal weighting is adopted. The target-side variance proxy is then defined as the average of these two estimates:

$$\mathrm{Var}_{t,j} = \frac{1}{2}\left[(\sigma_j^-)^2 + (\sigma_j^+)^2\right]. \tag{3-31}$$

This construction provides a simple approximation of the uncertainty implied by the Wilson interval while accounting for interval asymmetry. Wider Wilson intervals produce larger variance estimates, whereas tighter intervals imply greater certainty and therefore smaller target-side variance. The resulting quantity serves as a scale against which the concentration of the optimized prior can be evaluated.

Logarithmic variance-ratio regularization

The compatibility metrics introduced in Section 3-3 encourage the prior to place probability mass inside the feasible region. However, these terms alone do not prevent the optimizer from selecting arbitrarily large concentration parameters τ_j, producing overly concentrated priors that effectively collapse onto a small subset of $\mathcal{F}_i$. To discourage such behavior, a regularization term is introduced that compares the prior variance to the uncertainty implied by the target data.

For an optimized Beta prior parameterized by mean μ_j and concentration τ_j, the variance is

$$\mathrm{Var}_{p,j} = \frac{\mu_j(1-\mu_j)}{\tau_j + 1}. \tag{3-32}$$

The regularization term compares this variance against the target-side variance proxy through a logarithmic variance ratio:

$$\mathcal{R}_\tau(\boldsymbol{\mu}, \boldsymbol{\tau}) = \sum_{j=1}^{J}\left[\log\left(\frac{\mathrm{Var}_{p,j} + \varepsilon}{\mathrm{Var}_{t,j} + \varepsilon}\right)\right]^2 \tag{3-33}$$

where $\varepsilon > 0$ is a small constant introduced for numerical stability. The squared log-ratio is chosen because the logarithmic relative difference is the unique relative-difference measure that is simultaneously symmetric, additive, and scale-invariant [75]. First, the variance ratio provides a dimensionless and scale-invariant measure of uncertainty mismatch: a prior that is twice too concentrated is penalized equally regardless of the absolute magnitude of the variances. Second, the logarithmic transformation converts multiplicative deviations into additive ones, while squaring ensures symmetric penalization of both over-concentrated and over-diffuse priors, since $[\log(a/b)]^2 = [\log(b/a)]^2$. Finally, the penalty attains its unique minimum of zero when $\mathrm{Var}_{p,j} = \mathrm{Var}_{t,j}$, encouraging the optimized prior to remain broadly consistent with the uncertainty implied by the target data while still allowing the prior mean to adapt to the geometry of the feasible region.

3-6-5 Combined cost function

The full cost function combines the base cost function with the concentration regularization term:

$$\min_{\boldsymbol{\mu},\boldsymbol{\tau}} \mathcal{L}(\boldsymbol{\mu},\boldsymbol{\tau}) = \mathcal{L}_{\text{base}}(\boldsymbol{\mu},\boldsymbol{\tau}) + w_\tau \, \mathcal{R}_\tau(\boldsymbol{\mu},\boldsymbol{\tau}), \tag{3-34}$$

where $w_\tau > 0$ controls the strength of the concentration regularization. Substituting the definitions from Equations (3-29) and (3-33), the full cost function reads

$$\min_{\boldsymbol{\mu},\boldsymbol{\tau}} \left\{ w_v \, \mathbb{E}\big[V(\theta)\big] - w_m \, \mathbb{E}\big[\mathbf{1}[\theta \in \mathcal{F}_i]\big] + w_d \, \mathbb{E}\big[\|\theta - \theta_c\|_1\big] + w_\tau \sum_{j=1}^{J} \left[\log\left(\frac{\mathrm{Var}_{p,j} + \varepsilon}{\mathrm{Var}_{t,j} + \varepsilon} \right) \right]^2 \right\}, \tag{3-35}$$

where all expectations are taken over $\theta \sim \mathbb{P}_\phi$.

The violation term drives the prior mean toward the interior of the feasible region. The centering term discourages boundary solutions. The mass term rewards feasible probability mass. The regularization term prevents the optimizer from exploiting concentration as a shortcut, ensuring that the optimized prior reflects the genuine uncertainty supported by the target data. The entire optimization framework is illustrated in Figure 3-2.

The expectations in Equation (3-35) are evaluated by Monte Carlo sampling from the joint prior in Equation (3-26). Because the feasible region $\mathcal{F}_i$ couples all CPT rows through the cross-row constraints, the cost function introduces nonlinear interactions between rows even though the prior factorizes. The resulting optimization landscape is nonconvex and contains multiple local optima. The procedure used to solve this cost function is described in Section 3-7.

3-7 Optimization procedure

The cost function defined in Equation (3-35) is a nonconvex stochastic function of the hyperparameters $(\boldsymbol{\mu},\boldsymbol{\tau})$. Three properties determine the choice of optimization procedure:

1. The expectations are evaluated by Monte Carlo sampling and therefore cannot be differentiated analytically.
2. The prior mass term and the cross-row violation penalties introduce nonlinear interactions between rows through the joint feasibility check, preventing the cost function from decomposing into independent per-row subproblems.
3. The landscape contains multiple local optima due to the interplay between the violation, mass, and centering terms.

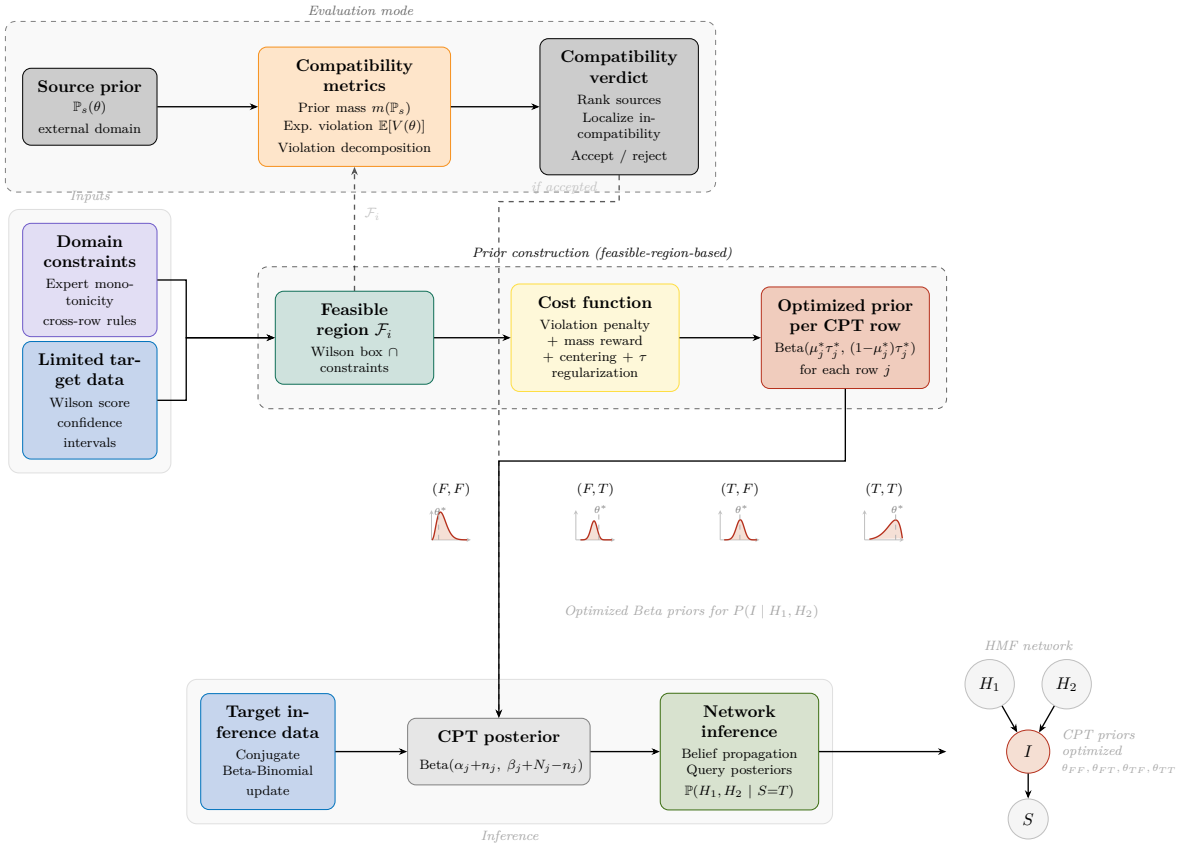


Figure 3-2: Overview of the proposed prior optimization framework. Domain constraints and limited target data are combined into the feasible region $\mathcal{F}_i$ (Section 3-2). Compatibility metrics are incorporated into a cost function minimized by Powell's method with multiple restarts (Sections 3-6–3-7), producing an optimized Beta prior for each row of the indicator CPT $\mathbb{P}(I \mid H_1, H_2)$. Dashed lines mark the true parameter values θ^*. The optimized priors are updated conjugately with target inference data to yield CPT posteriors, propagated via belief propagation to hypothesis query posteriors. The highlighted node I in the HMF network indicates where prior optimization is applied.

3-7-1 Monte Carlo approximation of the cost function

All three stochastic terms in Equation (3-35) are approximated by sample averages over S draws from the joint prior $p(\boldsymbol{\theta}; \boldsymbol{\mu}, \boldsymbol{\tau})$. This sample average approximation is a standard approach for stochastic optimization problems in which the cost function involves expectations that cannot be evaluated analytically [72, 76]. To ensure a deterministic cost landscape within each restart, the same random seed is used for all S draws on every function evaluation. This common-random-numbers technique makes the approximate cost function fully deterministic across iterations of a single run, so that the optimizer sees a fixed landscape rather than a noisy one.

3-7-2 Derivative-free optimization

The cost function is minimized using Powell's method, a derivative-free line-search algorithm that performs sequential one-dimensional minimizations along a set of conjugate directions without computing or approximating gradients [77]. Powell's method is well-suited to the present setting for three reasons. First, it requires only function evaluations, avoiding the instability that finite-difference gradient estimates introduce under Monte Carlo noise. Second, it is effective for moderate problem dimensions: the optimization variables are $(\boldsymbol{\mu}, \boldsymbol{\tau}) \in \mathbb{R}^{2J}$,

where J is the number of CPT rows, which remains small in the HMF networks considered here. Third, it imposes no convexity requirement on the cost function, making it applicable to the nonconvex landscape induced by the joint feasibility constraint.

To ensure that $\mu_j \in (0,1)$ and $\tau_j > 0$ throughout optimization, the variables are constrained to their admissible ranges via a smooth reparameterization. The specific implementation details are reported in Section 4-3.

As an alternative derivative-free method, Nelder–Mead simplex optimization [78] is also evaluated in the configuration study reported in Section 4-1. Nelder–Mead operates by maintaining and iteratively updating a simplex of $2J+1$ points in the parameter space and similarly requires only function evaluations. The comparative evaluation in Section 4-3 motivates the selection of Powell's method as the primary optimizer for all subsequent experiments.

3-7-3 Multi-start strategy

Because the cost function contains multiple local optima, a multi-start strategy is employed: the optimizer is run from R starting points and the solution with the lowest cost is retained:

$$(\boldsymbol{\mu}^*, \boldsymbol{\tau}^*) = \arg \min_{r=1,\dots,R} \mathcal{L}(\boldsymbol{\mu}^{(r)}, \boldsymbol{\tau}^{(r)}). \tag{3-36}$$

Starting points follow a mixed strategy. The first start is always the projected center of $\mathcal{F}_i$, providing a deterministic and geometrically motivated initialization. Subsequent starts alternate between perturbed-center draws and random draws from a wide fixed range $\mu_j \in (0.05, 0.95)$, with τ_j drawn uniformly from a moderate range $[\tau_{\min}, \tau_{\max}]$. This mixed strategy combines the stability of a deterministic center initialization with the diversity needed to escape local optima.

3-8 Inference with optimized priors

Once a prior over the CPT parameters has been selected or constructed, it is incorporated into standard conjugate Bayesian updating using the target inference data. The procedure is identical in both operating modes of the framework: the only difference lies in the origin of the prior hyperparameters (α_j, β_j).

In the construction mode, (α_j, β_j) are obtained from the solution $(\boldsymbol{\mu}^*, \boldsymbol{\tau}^*)$ of the optimization problem in Equation (3-34), via $\alpha_j = \mu_j^* \tau_j^*$ and $\beta_j = (1 - \mu_j^*)\tau_j^*$. In the evaluation mode, (α_j, β_j) are derived from the selected source prior $\mathbb{P}_s$ via method of moments, fitting a Beta distribution to its mean and variance. In both cases, the prior is updated conjugately with the target inference subset:

$$\theta_j \mid \mathcal{D}_{\text{inf}} \sim \text{Beta}(\alpha_j + n_j,\ \beta_j + N_j - n_j), \tag{3-37}$$

where n_j and N_j are the success count and total count for row j in the inference subset $\mathcal{D}_{\text{inf}}$. The posterior mean is used as the point estimate of θ_j:

$$\hat{\theta}_j^{\text{post}} = \frac{\alpha_j + n_j}{\alpha_j + \beta_j + N_j}. \tag{3-38}$$

The posterior CPT estimates are then used to parameterize the indicator node of the target BBN, while all other nodes are fixed at their specified parameter values. Network-level inference proceeds via standard belief propagation, propagating evidence from the statement layer upward through the indicator to the hypothesis nodes [25]. The resulting hypothesis posteriors constitute the operationally relevant output of the approach.

Parameter uncertainty in the posterior CPT propagates nonlinearly to network query posteriors through this inference chain, since belief propagation involves products of conditional probabilities and query posteriors are rational functions of the CPT parameters [67].

To characterize this propagation, CPT realizations are drawn from the row-wise Beta posteriors and propagated through the network, yielding a distribution over query posteriors. This second-order characterization is described in Section 3-9.

3-9 Validation design

This section defines the validation design used to evaluate the proposed framework. The design ensures a fair and statistically robust comparison across methods, with ground-truth parameters known throughout the simulation, enabling direct quantitative assessment of CPT posterior accuracy and downstream query inference quality.

3-9-1 Reproducibility

To ensure a fair comparison, all methods are evaluated under identical experimental conditions, including the same target datasets, the same calibration and inference splits, and a common set of evaluation metrics. The feasible region $\mathcal{F}_i$ is constructed from the same calibration data for all methods, and all priors are updated conjugately with the same inference subset.

To account for stochastic variability, all experiments are repeated over 15 independent random seeds. For each seed, a target dataset is drawn by forward sampling from the true BBN, and smaller target sizes are obtained as nested prefix subsets of this dataset. This reflects a realistic operational scenario in which a single dataset is available and method performance is evaluated across different levels of data availability. The nested design further ensures that performance differences across target sizes are not driven by unrelated sampling variation between seeds.

3-9-2 Evaluation metrics

Four quantitative metrics and one visualization method are used to evaluate and compare the methods. In the simulation setting, ground-truth parameters are known, enabling direct comparison of posterior estimates against the true CPT. In real-world applications, ground-truth parameters are unavailable and evaluation must rely on predictive performance and consistency with domain-informed constraints instead.

Each metric is computed at two levels: the CPT level, measuring the accuracy of the learned conditional probability table parameters, and the network level, measuring the accuracy of downstream hypothesis query posteriors. This parallel structure ensures that CPT-level improvements are evaluated not only in isolation but also in terms of their propagation to operationally relevant inference outputs.

Mean Absolute Error. Mean absolute error (MAE) is used at both the CPT and network level. At the CPT level, it measures the average absolute deviation between the posterior mean estimate and the true parameter value, averaged over rows:

$$\text{CPT MAE} = \frac{1}{J}\sum_{j=1}^{J}|\hat{\theta}_j^{\text{post}} - \theta_j^*|.$$

At the network level, it measures the average absolute deviation between query posteriors under the learned BBN and under the true BBN, averaged over a set of inference queries:

$$\text{Network MAE} = \frac{1}{Q}\sum_{q=1}^{Q}|\hat{p}_q - p_q^*|.$$

CPT MAE is the primary accuracy metric and measures point estimation quality. Network MAE evaluates whether CPT-level improvements propagate to downstream hypothesis inference, addressing Research Question **3C**:.

Root Mean Squared Error. Root mean squared error (RMSE) is computed at both levels:

$$\text{CPT RMSE} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(\hat{\theta}_j^{\text{post}} - \theta_j^*\right)^2}, \qquad \text{Network RMSE} = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{p}_q - p_q^*\right)^2}.$$

Compared to MAE, RMSE penalizes larger errors more heavily, making it more sensitive to seeds where prior misspecification causes a particularly poor estimate, at either the parameter or query level.

KL Divergence. KL divergence is computed at both levels. At the CPT level, the true parameter θ_j^* and the posterior mean $\hat{\theta}_j^{\text{post}}$ each define a Bernoulli distribution, and the divergence is averaged over rows:

$$\text{CPT KL} = \frac{1}{J}\sum_{j=1}^{J}\text{KL}\left(\theta_j^* \,\|\, \hat{\theta}_j^{\text{post}}\right).$$

At the network level, the true query posterior p_q^* and the estimated query posterior $\hat{p}_q$ each define a Bernoulli distribution, and the divergence is averaged over queries:

$$\text{Network KL} = \frac{1}{Q}\sum_{q=1}^{Q}\text{KL}\left(p_q^* \,\|\, \hat{p}_q\right).$$

Unlike MAE, KL divergence is sensitive to systematic displacement of the estimated value and penalizes errors near the probability boundaries more heavily. It is particularly informative for detecting source bias: a method may achieve low MAE while still exhibiting high KL divergence if estimates are displaced systematically in one direction. Because KL divergence is non-negative and right-skewed, results are reported using the 10th and 90th percentile across seeds rather than ± 1 standard deviation.

95% posterior CI coverage. The proportion of CPT rows for which the true parameter θ_j^* falls within the 95% credible interval of the Beta posterior. A well-calibrated prior should yield coverage close to 0.95.

Statistical significance. Paired one-sided Wilcoxon signed-rank tests are conducted to assess whether observed performance differences between methods are statistically reliable, following standard practice for non-parametric comparison of methods across repeated runs [79]. The full results are reported in Appendix B-3-3.

3-9-3 Query posterior uncertainty characterization

To characterize how CPT parameter uncertainty propagates to network-level query uncertainty, Monte Carlo realizations are drawn from the row-wise Beta posteriors and propagated through the BBN via belief propagation. For each method and seed, N_{MC} realizations are drawn independently, each parameterizing the indicator-node CPT. All network queries are evaluated under each realization, yielding a sample from the induced posterior distribution over query probabilities.

For visualization purposes, query posterior distributions are represented using Gaussian kernel density estimation (KDE) [80]. As a non-parametric density estimator, KDE provides a flexible representation of propagated uncertainty without assuming a specific parametric form for the resulting query distributions. A fixed absolute bandwidth is used to ensure comparable smoothing across methods with different uncertainty levels. The specific values of N_{MC} and the evaluated queries are reported in Chapter 4.

3-10 Summary

This chapter presented the methodological framework for feasible-region-based Bayesian prior optimization in BBNs under data scarcity, addressing Research Questions 1, 2, and 3A.

The framework begins with the construction of a feasible parameter region $\mathcal{F}_i$, defined as the intersection of data-driven Wilson score uncertainty intervals and qualitative cross-row constraints encoding domain knowledge (RQ1). This region makes explicit which parameter configurations are simultaneously consistent with the observed target data and the structural assumptions of the domain.

Two compatibility metrics evaluate how well a candidate prior aligns with $\mathcal{F}_i$: prior mass, measuring the total probability assigned to admissible configurations, and expected constraint violation, quantifying the severity of infeasibility for configurations outside the region (RQ2A). These metrics first serve in an evaluation mode, where they assess and rank candidate source priors against the feasible region. By evaluating compatibility jointly over the feasible region rather than independently per row, the framework provides the basis for comparing joint and row-wise assessments in Chapter 4 (RQ2B). A violation decomposition further attributes global incompatibility to specific rows and cross-row constraints, providing a structured basis for targeted refinement (RQ2C).

The same metrics are incorporated into a cost function over a parametric Beta prior family (RQ3A). The cost function combines three terms: an expected violation penalty that drives the prior mean toward the interior of $\mathcal{F}_i$, a prior mass reward that directly incentivizes feasible probability mass, and a centering term based on the distance to the representative interior of $\mathcal{F}_i$, which discourages fragile boundary solutions. An target-uncertainty regularization term is added to prevent overconfident priors by penalizing mismatches between the prior variance and a target-side variance proxy derived from the Wilson interval widths. This construction mode requires no source domain data: a structured, feasibility-consistent prior is built entirely from target-side information.

The resulting nonconvex stochastic cost function is minimized using Powell's derivative-free method with a multi-start strategy, where starting points combine a systematic initialization at the feasible-region center with random restarts to escape local optima. The optimized prior is then updated conjugately with target inference data, and the posterior CPT estimates are propagated through the BBN via belief propagation to yield operationally relevant hypothesis posteriors. To characterize how parameter uncertainty propagates to query-level outputs, Monte

Carlo realizations are drawn from the posterior and propagated through the network, yielding a distribution over query posteriors, which is visualized using Gaussian kernel density estimation (KDE).

The experimental evaluation of this framework, including validation of the compatibility metrics, comparison between joint and row-wise compatibility assessment, benchmark comparison of the proposed method with MPL-TC, and network-level inference results, is reported in Chapter 4.

Chapter 4

Experimental setup and case study

This chapter evaluates the proposed prior optimization framework through a controlled simulation study, addressing Research Questions 2 and 3. Ground-truth CPT parameters are known throughout, enabling direct quantitative evaluation of posterior accuracy and downstream inference quality across varying target sample sizes and constraint tightness levels.

The chapter is organized as follows. Section 4-1 describes the simulated BBN, the experimental design, the comparison baselines, and the evaluation metrics. Section 4-2 validates the compatibility metrics against source priors with known levels of mismatch and evaluates incompatibility localization via violation decomposition, addressing RQ2 and subquestions. Section 4-3 evaluates prior optimization and CPT parameter-learning performance, including benchmark comparison with MPL-TC, addressing RQ3B. Section 4-4 reports network-level inference results, evaluating whether CPT-level improvements propagate to downstream hypothesis posteriors, addressing RQ3C.

4-1 Experimental setup

This section describes the simulation environment used to evaluate the proposed framework. A small four-node BBN following the HMF hierarchy is used as the target model, with known ground-truth CPT parameters that enable direct quantitative comparison of posterior accuracy across methods. The setup covers the simulated network and target model, the feasible-region construction procedure, the methods compared in the parameter-learning benchmark, the source prior families used for compatibility metric validation and the MPL-TC baseline, the evaluation metrics, and the full factorial experimental design.

4-1-1 Simulated BBN and target model

The simulation experiment is conducted on a small four-node BBN following the HMF hierarchy introduced in Chapter 2. The network contains two binary hypothesis nodes H_1 and H_2, one binary indicator node I, and one binary observation node S, with the dependency structure shown in Figure 4-1.

The primary object of analysis is the CPT $\mathbb{P}(I \mid H_1, H_2)$, which contains four rows corresponding to the parent configurations $(H_1, H_2) \in \{(F,F),(F,T),(T,F),(T,T)\}$, denoted θ_{FF}, θ_{FT},

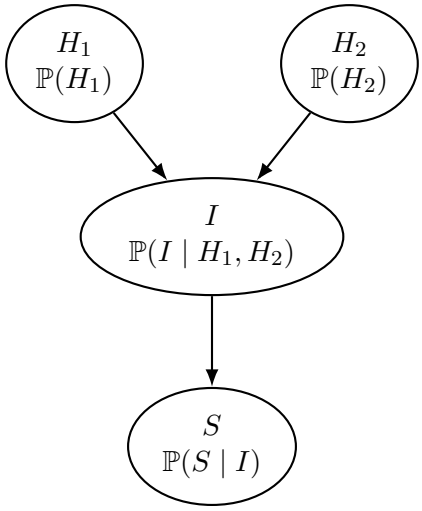


Figure 4-1: Simulated HMF network used throughout the simulation experiment. The primary object of analysis is the CPT $\mathbb{P}(I \mid H_1, H_2)$, consisting of four rows corresponding to the four binary parent configurations.

θ_{TF}, and θ_{TT} respectively. Parameterizing the indicator as a full joint table over both hypotheses, rather than as two independent marginal CPTs, relaxes the no-synergy assumption of the standard HMF (Section 2-6). At this scale the joint table has only four rows, so this introduces no additional elicitation burden relative to two marginal CPTs, while allowing synergistic effects between hypotheses to be represented and making the cross-row constraints C_1-C_7 fully expressive. The ground-truth parameter values are given in Table 4-1.

Table 4-1: Ground-truth CPT parameters for $\mathbb{P}(I = \text{Present} \mid H_1, H_2)$ in the simulated target model.

H_1	H_2	θ^*
False	False	0.10
False	True	0.70
True	False	0.60
True	True	0.90

These values encode two monotonicity relationships that serve as the basis for the cross-row constraints used throughout the experiments: the indicator probability increases when H_1 is active compared to when it is inactive, for any fixed state of H_2; and the indicator probability increases when H_2 is active compared to when it is inactive, for any fixed state of H_1. These relationships reflect the domain assumption that each hypothesis, when active, makes the indicator more likely to be present regardless of the other hypothesis's state. The prior probabilities on the root nodes are fixed at $\mathbb{P}(H_1 = T) = 0.3$ and $\mathbb{P}(H_2 = T) = 0.4$. The observation node CPT is fixed at $\mathbb{P}(S = T \mid I = \text{Present}) = 0.8$ and $\mathbb{P}(S = T \mid I = \text{Absent}) = 0.05$.

Synthetic target datasets are generated by forward sampling from this BBN. For each random seed, a single parent dataset of size $N_{\max} = 200$ is drawn first. Datasets at smaller target sizes are obtained as prefix subsets: the first N_{target} rows are used for each value of N_{target}. This nested design ensures that performance differences across data sizes are not driven by unrelated sampling variation between seeds, and realistically simulates the situation where an analyst progressively acquires additional observations. A consequence of this design is that results at different N_{target} values within the same seed are not statistically independent, which is accounted for in the significance analysis described in Section 4-1-6.

4-1-2 Feasible-region construction

For each seed and each N_{target}, the available target dataset is split 50/50 into a calibration subset and an inference subset. The calibration subset is used to construct Wilson score intervals and define the target-side feasible region. The inference subset is withheld until after prior construction and is used solely for posterior updating and evaluation. This separation ensures that the optimized prior is not evaluated on the same data used to construct the feasible region.

From the calibration subset, row-wise empirical proportions $\hat{p}_j = n_j/N_j$ are computed for each parent configuration j, where N_j is the number of calibration observations with parent configuration j and n_j is the number of those observations in which I = Present. Wilson score intervals at the 95% confidence level are computed for each row, yielding a box-shaped uncertainty set $\mathcal{B}_i$.

Throughout all experiments, 95% Wilson score intervals are used, corresponding to a standard normal critical value $z = 1.96$. This confidence level was selected because it provides a commonly used compromise between interval width and uncertainty quantification and is therefore used consistently across all experimental conditions.

Two constraint modes are evaluated, corresponding to different cross-row constraint sets imposed on the CPT parameters $(\theta_{FF}, \theta_{FT}, \theta_{TF}, \theta_{TT})$:

- **Mild constraints** ($\varepsilon = 0.30$): four hard monotonicity constraints and two closeness constraints with tolerance $\varepsilon = 0.30$.
- **Strict constraints** ($\varepsilon = 0.15$): the same six constraints as mild, with tighter tolerance $\varepsilon = 0.15$, plus one additional gap constraint.

The two tolerance values were selected to create two qualitatively different feasible-region geometries.

The constraints imposed under both modes are:

$$C_1 : \theta_{FF} \leq \theta_{FT} \quad \text{(indicator more likely when } H_2 \text{ active, } H_1 \text{ inactive)}, \tag{4-1}$$
$$C_2 : \theta_{FF} \leq \theta_{TF} \quad \text{(indicator more likely when } H_1 \text{ active, } H_2 \text{ inactive)}, \tag{4-2}$$
$$C_3 : \theta_{FT} \leq \theta_{TT} \quad \text{(activating } H_1 \text{ cannot reduce indicator when } H_2 \text{ is active)}, \tag{4-3}$$
$$C_4 : \theta_{TF} \leq \theta_{TT} \quad \text{(activating } H_2 \text{ cannot reduce indicator when } H_1 \text{ is active)}, \tag{4-4}$$
$$C_5 : \theta_{FT} \leq \theta_{TF} + \varepsilon \quad \text{(cross-term closeness, one direction)}, \tag{4-5}$$
$$C_6 : \theta_{TF} \leq \theta_{FT} + \varepsilon \quad \text{(cross-term closeness, other direction)}. \tag{4-6}$$

Constraints C_1-C_4 are hard monotonicity constraints that encode the domain assumption that the indicator is more likely to be present when either hypothesis is active. These hold without tolerance. Constraints C_5 and C_6 encode the assumption that neither hypothesis dominates the other in isolation: absent domain knowledge specifying which hypothesis has a stronger marginal effect on the indicator, the indicator probabilities under single-hypothesis activation should not differ by more than ε. This prevents the feasible region from collapsing to a single ordering of θ_{FT} and θ_{TF}, preserving uncertainty about which single-hypothesis condition is more diagnostic. The separation of C_5 and C_6 allows their contributions to the total violation score to be analyzed individually.

Under strict constraints, one additional gap constraint is imposed:

$$C_7 : \theta_{TT} \geq \theta_{FF} + 0.30, \tag{4-7}$$

encoding the assumption that the joint activation of both hypotheses must produce a meaningfully higher indicator probability than joint inactivation. The gap of 0.30 is chosen to be consistent with the true parameter difference $\theta^*_{TT} - \theta^*_{FF} = 0.90 - 0.10 = 0.80$, ensuring that the constraint is satisfied by the ground-truth CPT with substantial margin.

The joint feasible region is then defined as:

$$\mathcal{F}_{i,\text{mode}} = \{\theta \in \mathcal{B}_i \mid g_\ell(\theta) \leq 0 \quad \forall\, \ell = 1, \ldots, K\}\,,$$

where $K = 6$ under mild constraints ($\mathcal{C}_{\text{mild}}$) and $K = 7$ under strict constraints ($\mathcal{C}_{\text{strict}}$).

The subscript i denotes the target CPT node under analysis, while the additional mode index indicates which constraint set is used.

Concretely, the constraint functions are:

$$\begin{aligned}
g_1(\theta) &= \theta_{FF} - \theta_{FT},\\
g_2(\theta) &= \theta_{FF} - \theta_{TF},\\
g_3(\theta) &= \theta_{FT} - \theta_{TT},\\
g_4(\theta) &= \theta_{TF} - \theta_{TT},\\
g_5(\theta) &= \theta_{FT} - \theta_{TF} - \varepsilon,\\
g_6(\theta) &= \theta_{TF} - \theta_{FT} - \varepsilon,\\
g_7(\theta) &= 0.30 - (\theta_{TT} - \theta_{FF}) \quad \text{(strict only)}.
\end{aligned}$$

4-1-3 Experimental modes

The simulation study addresses two distinct evaluation goals, which require different experimental setups and comparison methods. These are organized into two modes that are kept separate throughout the chapter.

Evaluation mode validates whether the compatibility metrics correctly rank source priors of known and increasing mismatch against the feasible region. No prior optimization is performed in this mode; the metrics are applied to fixed source prior families and evaluated for ranking consistency, sensitivity to constraint tightness, and stability across seeds.

Optimization mode evaluates prior construction and downstream inference performance in the setting where no source prior is available. A feasible-region-based prior is constructed entirely from the target calibration data and qualitative constraints, and compared against a uniform prior baseline and MPL-TC transfer learning baselines of varying source quality.

The source prior families used in Mode 1 and the MPL-TC source datasets used in Mode 2 are independent of each other and serve different experimental purposes, as described in the following two subsections.

4-1-4 Evaluation mode: Source prior families for compatibility metric validation

To evaluate whether the compatibility metrics correctly distinguish well-aligned from poorly-aligned source priors, six source prior families are constructed with controlled and increasing deviation from the target CPT. Each family generates a row-wise Beta prior by fixing the prior means to shifted versions of the true target parameters. All six families are assigned the same equivalent sample size τ, so that only the prior location varies across families. This isolates the effect of prior-mean mismatch and ensures that differences in compatibility scores reflect alignment with the feasible region rather than differences in prior concentration.

Table 4-2: Source prior families used for compatibility metric validation, ordered by increasing mismatch from the target. Values give $\mathbb{P}(I = \text{Present} \mid H_1, H_2)$ for each parent configuration. Target values: $(\theta_{FF}, \theta_{FT}, \theta_{TF}, \theta_{TT}) = (0.10,\ 0.70,\ 0.60,\ 0.90)$.

Source family	Description	θ_{FF}	θ_{FT}	θ_{TF}	θ_{TT}
Equal	Identical to target; upper bound on compatibility.	0.10	0.70	0.60	0.90
Very mild	Very small deviations; all values within target Wilson intervals.	0.12	0.68	0.58	0.88
Mild	Small shift; monotonicity preserved but some rows outside Wilson intervals.	0.15	0.65	0.55	0.80
Moderate*	Larger shift; partial violations of the feasible region. Also shifts H_1, H_2 priors and S CPT.	0.20	0.75	0.50	0.85
Strong shift	Large deviations; most rows outside the credal box.	0.30	0.50	0.45	0.65
Conflict	Deliberately violates both monotonicity constraints; indicator probability decreases as hypotheses become active.	0.80	0.60	0.70	0.50

*The moderate source additionally sets $\mathbb{P}(H_1{=}T) = 0.50$, $\mathbb{P}(H_2{=}T) = 0.30$, $\mathbb{P}(S{=}T \mid I{=}\text{Absent}) = 0.10$, and $\mathbb{P}(S{=}T \mid I{=}\text{Present}) = 0.70$.

The six families are ordered by increasing mismatch and summarized in Table 4-2.

These families are used exclusively in Section 4-2 and do not appear in the parameter learning or network inference experiments.

4-1-5 Optimization mode: Methods compared for prior optimization and inference

Three approaches to prior specification are evaluated and compared against the ground-truth CPT in Mode 2.

- **Uniform prior.** A Beta$(1, 1)$ prior is placed on each CPT row independently, representing complete prior ignorance. This serves as the baseline against which all other methods are compared.
- **Optimized prior.** Row-wise Beta priors are constructed by minimizing the cost function defined in Chapter 3, using the calibration-split Wilson intervals and cross-row constraints to define the feasible region $\mathcal{F}_{i,\text{mode}}$. Powell's method with 10 random restarts is used; the solution with the lowest cost across restarts is selected. Cost function weights are fixed at $w_v = 5.0$, $w_m = 1.0$, $w_d = 0.25$, and $w_\tau = 0.02$, selected based on the configuration validation reported in Section 4-3-1. No source domain data is used at any stage.
- **MPL-TC baseline.** A simplified implementation of the MPL-TC method [8] is used as a transfer learning comparator. For each CPT row, the source dataset is bootstrapped to estimate the sampling distribution of the maximum-likelihood parameter estimate. A Beta prior is then fitted through moment matching to the bootstrap mean and variance, yielding row-specific concentration parameters that reflect source-domain uncertainty. The resulting prior is updated conjugately with the target inference data.

Both the optimized prior and the MPL-TC baseline are updated conjugately with the inference subset using Beta-Binomial updating after prior construction. No cross-row constraints are applied during MPL-TC prior construction: each row is treated independently from source data. The proposed method, by contrast, optimizes all rows jointly against the feasible region $\mathcal{F}_{i,\text{mode}}$. This makes the comparison conservative: any observed improvement of the optimized method over MPL-TC represents a lower bound on the benefit of feasible-region-based prior construction.

Three source quality levels are used for the MPL-TC baseline, generated from independently drawn source datasets. These correspond to the mild, moderate, and conflict settings from Table 4-2, adapted for use with bootstrap-based prior fitting rather than fixed prior means. Their parameter values are reproduced in Table 4-3 for reference.

Table 4-3: MPL-TC source quality levels used in the parameter-learning benchmark. Values give $\mathbb{P}(I = \text{Present} \mid H_1, H_2)$ for each parent configuration. Target: (0.10, 0.70, 0.60, 0.90).

Source	Description	θ_{FF}	θ_{FT}	θ_{TF}	θ_{TT}
Mild	Small shift; monotonicity preserved. Moderately related domain.	0.15	0.65	0.55	0.80
Moderate*	Larger shift; partial feasible-region violations. Partially related domain.	0.20	0.75	0.50	0.85
Conflict	Violates both monotonicity constraints; misleading source domain.	0.80	0.60	0.70	0.50

*Also shifts $\mathbb{P}(H_1{=}T) = 0.50$, $P(H_2{=}T) = 0.30$, $\mathbb{P}(S{=}T \mid I{=}\text{Absent}) = 0.10$, $\mathbb{P}(S{=}T \mid I{=}\text{Present}) = 0.70$.

4-1-6 Experimental design and performance metrics

Results are assessed using CPT-level and network-level MAE, RMSE, and Bernoulli KL divergence, with calibration evaluated through 95% posterior credible-interval coverage and statistical significance assessed using paired one-sided Wilcoxon signed-rank tests. Query posterior distributions are characterized via KDE over Monte Carlo samples. Full metric definitions are given in Section 3-9-2.

Table 4-4 summarizes the full factorial design of the simulation experiment.

Table 4-4: Factorial design of the simulation experiment. Each combination of seed, N_{target}, and constraint mode constitutes one experimental run.

Factor	Values
N_{target}	10, 20, 30, 50, 75, 100, 150, 200
Degree of mismatch of $\mathbb{P}_s$	Equal, very mild, mild, moderate, strong shift, conflict
Constraint mode	Mild ($\varepsilon = 0.30$), Strict ($\varepsilon = 0.15$)
Methods	Uniform prior, optimized prior, MPL-TC (mild, moderate, conflict source)
Seeds	20 (metric validation) / 15 (optimization experiments); nested prefix design, $N_{\max} = 200$
Calibration split	50% calibration, 50% inference
Optimization method	Powell, multi-start with 10 restarts per run
MPL-TC source size	Equal to calibration split ($N_{\text{src}} = N_{\text{target}}/2$); sensitivity over $N_{\text{src}} \in \{25, 50, 100, 200\}$

For Mode 2, the following pipeline is executed for each combination of seed, N_{target}, and constraint mode:

1. The target prefix of size N_{target} is split 50/50 into calibration and inference subsets.
2. Wilson score intervals are computed from the calibration subset to define the credal box $\mathcal{B}_i$.
3. The feasible-region-based prior is constructed by minimizing the cost function over $\mathcal{F}_{i,\text{mode}}$ using 10 Powell restarts. No source domain data is used at this stage.
4. For comparison, MPL-TC priors are constructed from an independently generated source dataset. The source dataset contains the same number of observations as the target calibration split ($N_{\text{src}} = N_{\text{target}}/2$), ensuring a fair comparison under equal data availability. A sensitivity analysis varying $N_{\text{src}} \in \{25, 50, 100, 200\}$ is reported in Section 4-3-3. The source families used for compatibility validation are independent of the source datasets used for the MPL-TC benchmark. The latter use only the mild, moderate and conflict settings.

5. All priors are updated conjugately with the inference subset using Beta-Binomial updating.

6. All evaluation metrics are computed against the ground-truth CPT parameters and the true BBN query posteriors.

Because dataset prefixes at different N_{target} values share observations within the same seed, results across the N axis should be interpreted as a performance trajectory rather than as a sequence of independent hypothesis tests. Full significance test results are reported in Appendix B-3-3.

4-2 Compatibility metric validation

This section evaluates whether the proposed compatibility metrics behave as intended when applied to source priors with known mismatch levels. This validation step is necessary: if the metrics cannot distinguish well-aligned from poorly aligned priors, the cost function built upon them lacks a meaningful foundation.

The evaluation focuses on three properties:

1. **Ranking:** the ability to produce the expected ordering of source families, from equal to conflict.

2. **Sensitivity:** the extent to which discriminative power increases as the feasible region tightens through additional target data or stricter constraints.

3. **Robustness:** the stability of the ranking across different target datasets and random seeds.

4-2-1 Evaluation procedure

The compatibility metric validation follows Mode 1 of the experimental design described in Section 4-1-6. The six source families from Table 4-2 are evaluated against feasible regions constructed from target datasets of varying size under both constraint modes. For each source family and target size, 10,000 samples are drawn from the source prior and evaluated against the feasible region $\mathcal{F}_{i,\text{mode}}$.

The expected compatibility ranking is summarized in Table 4-5.

Table 4-5: Expected ranking of source families by compatibility metric, from most compatible (rank 1) to least compatible (rank 6).

Rank	Source family	Prior mass ↑	Expected violation ↓
1	Equal	highest	lowest
2	Very mild		
3	Mild		
4	Moderate		
5	Strong shift		
6	Conflict	lowest	highest

4-2-2 Row-wise vs joint compatibility

Compatibility can be assessed at two levels of granularity. Row-wise assessment evaluates each CPT row independently against its corresponding Wilson interval, without considering relationships between rows. Joint assessment evaluates all rows simultaneously against the full feasible region $\mathcal{F}_{i,\text{mode}}$, which incorporates both the interval constraints and the cross-row monotonicity constraints.

Row-wise metrics provide interpretable local diagnostics: they identify which individual rows are most misaligned with the target data. However, they cannot detect incompatibilities that arise only at the level of cross-row constraints. For example, a source prior may have each row within its Wilson interval individually while still violating a monotonicity constraint between rows. The joint metric captures this structural mismatch while row-wise assessment misses it entirely.

The validation therefore considers both row-wise and joint compatibility metrics. The latter incorporates the full feasible-region structure and is expected to better capture incompatibilities arising from cross-row constraints. Row-wise diagnostics remain useful for the incompatibility localization described in Section 4-2-5, which decomposes global incompatibility into row-level and cross-row-level contributions to guide targeted intervention.

4-2-3 Ranking and sensitivity to constraint tightness

Table 4-6 reports the full trajectory of both metrics across all target sizes under the strict constraint mode.

Table 4-6: Prior mass in the feasible region and mean expected violation for each source family across all target sizes, under strict constraints ($\varepsilon = 0.15$), averaged over 20 seeds. The rows ordered by prior mass at $N = 75$, under which the strong-shift source ranks above the moderate source at small to moderate target sizes despite its larger nominal deviation, because the moderate source places more mass on constraint-violating configurations and is penalized more heavily by the feasible-region geometry.

	Prior mass ↑								Mean violation ↓							
Source	10	20	30	50	75	100	150	200	10	20	30	50	75	100	150	200
Equal	0.300	0.273	0.244	0.211	0.161	0.144	0.101	0.076	0.110	0.118	0.134	0.139	0.151	0.158	0.178	0.192
Very mild	0.271	0.245	0.220	0.191	0.150	0.131	0.092	0.070	0.129	0.139	0.156	0.161	0.175	0.181	0.206	0.222
Mild	0.129	0.114	0.105	0.091	0.069	0.057	0.036	0.027	0.252	0.268	0.289	0.296	0.320	0.329	0.376	0.402
Strong shift	0.253	0.201	0.163	0.115	0.069	0.050	0.026	0.016	0.169	0.203	0.233	0.278	0.330	0.350	0.409	0.441
Moderate	0.108	0.097	0.079	0.063	0.043	0.037	0.027	0.020	0.264	0.288	0.323	0.330	0.354	0.362	0.394	0.418
Conflict	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.628	1.770	1.881	2.045	2.161	2.227	2.320	2.380

Figure 4-2 shows the prior mass in the feasible region for each source family as a function of target sample size, across three constraint regimes: no cross-row constraints, mild constraints ($\varepsilon = 0.30$), and strict constraints ($\varepsilon = 0.15$). Two patterns are immediately visible.

First, the metrics produce a broadly plausible ordering across all constraint regimes. The equal source consistently achieves the highest prior mass, and the conflict source is consistently ranked lowest at all target sizes and constraint levels. This confirms that the metric can detect severe cross-row ordering violation even under severe data scarcity, when Wilson intervals are wide and the feasible region is weakly constrained.

Second, discriminative power increases substantially as constraints tighten. Under no constraints, the feasible region reduces to the credal box $\mathcal{B}_i$ defined by the Wilson intervals alone. In this setting, the equal and very mild sources are nearly indistinguishable, and the intermediate families (mild, moderate, strong) overlap considerably at small N. Once cross-row monotonicity constraints are introduced, sources that appeared near-equivalent separate clearly: the mild

and moderate families drop to substantially lower prior mass, while the conflict source reaches near-zero mass under both constraint modes, making it trivially identifiable regardless of target sample size.

This finding has a direct implication for the optimization approach: domain-informed qualitative knowledge, encoded through cross-row constraints, makes the cost function more informative and therefore provides a stronger signal for prior construction. The no-constraint setting effectively reduces the framework to interval-based regularization alone, which is a meaningful but weaker baseline.

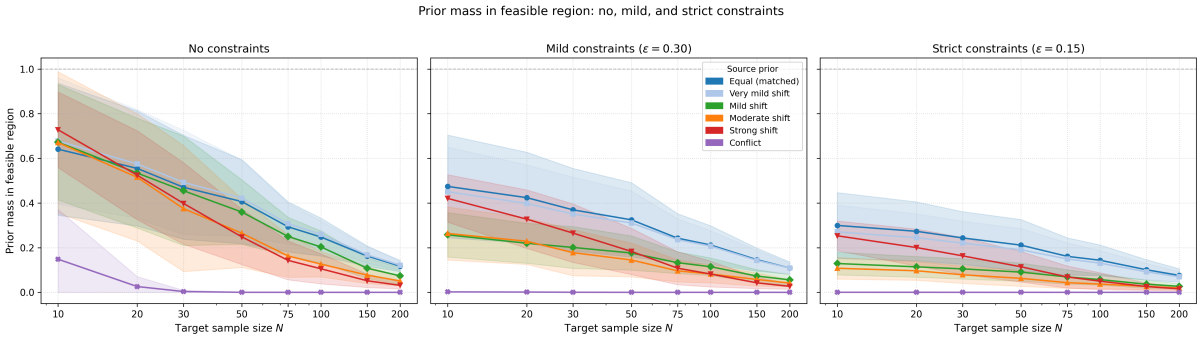


Figure 4-2: Prior mass in the feasible region as a function of target sample size for each source family, under no constraints (left), mild constraints (center), and strict constraints (right). Higher values indicate better compatibility with the target-side feasible region. The conflict source achieves near-zero mass under both constraint modes. Discriminative power between intermediate families increases substantially as constraints tighten. Shaded bands indicate ± 1 standard deviation across seeds.

Figure 4-3 reports the complementary metric: mean expected violation of the feasible-region constraints. The violation metric confirms the ranking established by prior mass from a different angle. The conflict source exhibits the highest violation at all target sizes and constraint levels, while the equal and very mild sources show near-zero violation throughout. Crucially, the conflict source is jointly identifiable by both metrics simultaneously: it achieves near-zero prior mass and maximum expected violation, making detection robust to the choice of metric.

A notable pattern in the violation plots is that even sources that appear moderately compatible by prior mass can show non-trivial violation under strict constraints. This occurs when a source prior places mass near the boundary of the feasible region: technically feasible samples exist, but much of the prior density is concentrated in regions that violate the monotonicity ordering. The joint evaluation of mass and violations therefore is more informative than either metric alone.

A further notable pattern visible in the violation columns of Table 4-6 is that the mean violation of the equal source increases monotonically with N_{target}, from 0.110 at $N = 10$ to 0.192 at $N = 200$ under strict constraints. This is counterintuitive at first glance, the equal source is identical to the target model, so one might expect near-zero violation throughout. The explanation lies in the structure of the feasible region $\mathcal{F}_{i,\text{mode}}$: as N_{target} increases, the Wilson intervals that define the credal box $\mathcal{B}_i$ narrow around the empirical proportions. The equal source prior is a continuous Beta distribution centered on the true parameter values rather than a point mass, so a growing fraction of its prior density falls outside the shrinking box $\mathcal{B}_i$, and therefore outside $\mathcal{F}_{i,\text{mode}}$, even though its mean is perfectly aligned with the target. This reflects a fundamental property of the framework: as more target data become available, the feasible region becomes more selective, and even a well-aligned continuous prior will inevitably place some mass outside it. The violation metric therefore measures not just directional mismatch but also the concentration of the source prior relative to the data-driven uncertainty set. This

property becomes relevant in the optimization context in Section 4-3: a prior that is directionally correct but overly diffuse will still incur non-trivial violation, motivating the τ-regularization term in the cost function.

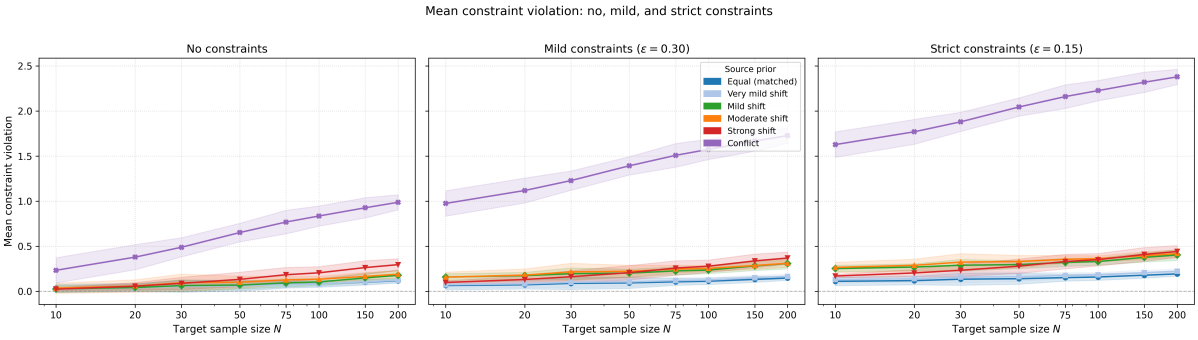


Figure 4-3: Mean expected violation of the feasible region as a function of target sample size for each source family, under no constraints (left), mild constraints (center), and strict constraints (right). Lower values indicate better consistency with the target-side constraints. The conflict source exhibits consistently high violation across all settings. Shaded bands indicate ± 1 standard deviation across seeds.

4-2-4 Robustness across seeds

The stability of the rankings among the seeds is assessed by examining whether the extreme families maintain their positions in all 20 random seeds. The full per-seed ranking stability is shown in Figure B-3 (Appendix B-1-2). The equal source is top-ranked and the conflict source is bottom-ranked by prior mass in every seed repetition, under all constraint modes and all target sizes. This confirms that the conflict source is consistently detected across all experimental conditions.

The relative ordering of the intermediate families shows more variability, particularly at small N where wide Wilson intervals produce a large feasible region with low discriminative power. At $N = 10$, the mild, moderate, and strong shift sources can produce overlapping prior mass estimates across seeds, making their relative ranking unreliable. This variability decreases monotonically as N increases: by $N = 100$ the ordering is stable in the majority of seeds under both constraint modes.

This pattern reflects a fundamental property of the framework: the feasible region defines a set of acceptable priors rather than a unique optimum. When data is scarce, multiple source families are consistent with available evidence and cannot be reliably distinguished on the basis of compatibility alone. This supports the optimization approach in Section 4-3, which constructs a structured prior within the feasible region rather than selecting from a discrete set of source candidates.

4-2-5 Incompatibility localization

The compatibility metrics produce a scalar score for a source prior as a whole. However, the violation decomposition $g_\ell(\theta) \leq 0$ is defined per constraint, and each constraint involves specific CPT rows. The global violation score is therefore decomposable into row-level and cross-row-level contributions, identifying which rows and cross-row ordering relationships of a source prior are responsible for incompatibility with the target feasible region.

Figures 4-4 and 4-5 show the row-wise and constraint-level violation profiles for four source families at $N = 50$, strict constraints, averaged over 20 seeds.

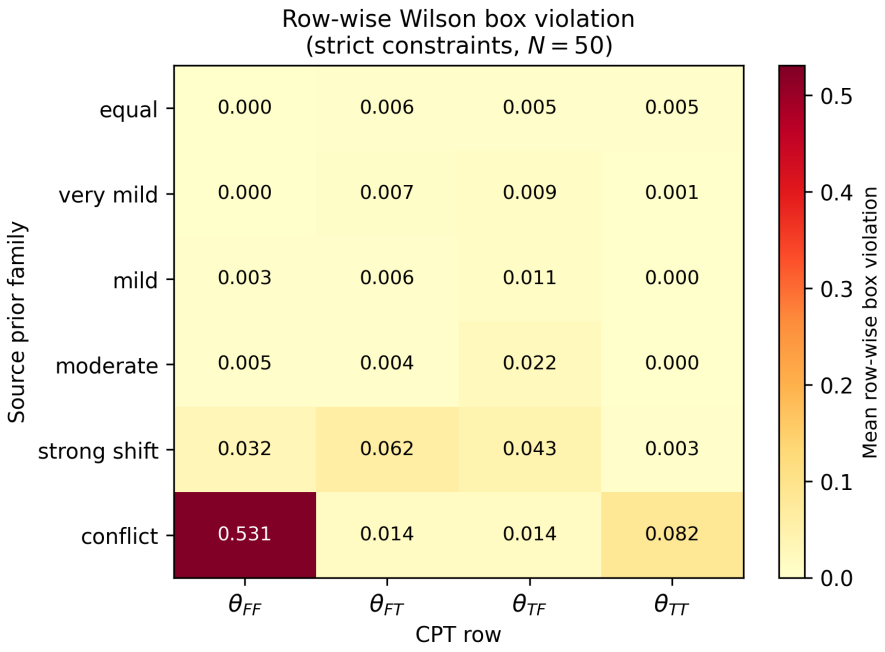


Figure 4-4: Row-wise box violation heatmap for all source families at $N = 50$, strict constraints, averaged over 20 seeds. Each cell reports the mean violation of that CPT row against its Wilson interval bound. The conflict source concentrates violation almost entirely on θ_{FF} (0.531), indicating that its main row-wise mismatch is the inflated baseline probability assigned to the (F, F) parent configuration. The remaining rows show substantially smaller violations. The equal, very mild, and mild shift sources show near-zero row-wise violation, while the strong shift source produces a moderate but more evenly distributed violation profile.

Two localization properties are visible in the figures. First, the row-wise heatmap shows that the conflict source is not diffusely incompatible with the target-side intervals. Instead, its box-level violation is concentrated almost entirely on θ_{FF}, the row with the largest absolute deviation from the target CPT. This correctly identifies the inflated baseline probability associated with the (F, F) parent configuration as the dominant row-wise source of incompatibility.

Second, the constraint-level decomposition provides information that is not visible from row-wise box violation alone. Although the conflict source shows only limited row-wise violation outside θ_{FF}, it violates several cross-row relationships because the inflated θ_{FF} changes the ordering of the CPT rows. Conversely, the equal source shows near-zero row-wise violation but can still violate a closeness constraint when θ_{FT} and θ_{TF} fall near the boundary of the allowed tolerance band. This illustrates why row-wise interval diagnostics and joint feasible-region evaluation are complementary: the former localizes marginal disagreement with the data-driven intervals, while the latter identifies structural incompatibility induced by cross-row constraints. When a candidate source prior exists and must be evaluated before use, the decomposition provides a structured basis for targeted expert review: an analyst can identify precisely which rows and constraints drive incompatibility rather than relying on a single aggregate score.

4-3 Prior optimization and Parameter Learning

The previous section showed that compatibility metrics based on feasible-region mass and expected violation reliably distinguish well-aligned from poorly aligned priors. The same metrics are now used constructively: rather than evaluating existing priors, they are incorporated into a cost function that constructs a prior directly from target-side information. The prior is pa-

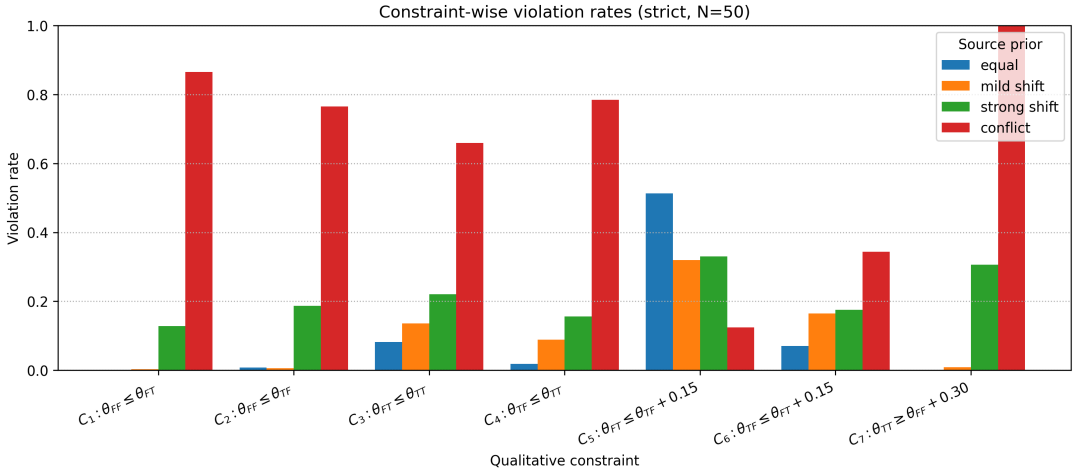


Figure 4-5: Constraint-level violation rates for four source families at $N = 50$, strict constraints. The conflict source (red) exhibits the highest violation rates across most constraints, with complete violation of the gap constraint C_7 ($\theta_{TT} \geq \theta_{FF} + 0.30$), followed by substantial violations of the monotonicity constraints C_1 ($\theta_{FF} \leq \theta_{FT}$), C_4 ($\theta_{TF} \leq \theta_{TT}$), and C_2 ($\theta_{FF} \leq \theta_{TF}$). These constraints are directly affected by the inflated θ_{FF} value identified in the row-wise heatmap. The equal source shows a notable violation on the closeness constraint C_5 ($\theta_{FT} \leq \theta_{TF} + 0.15$), indicating that θ_{FT} and θ_{TF} lie near the boundary of the allowed tolerance band even when individual rows remain consistent with their Wilson intervals.

rameterized row-wise as a Beta distribution in (μ, τ) form, where μ represents the prior mean and τ the equivalent sample size. The full cost function is defined in Equation 3-34.

4-3-1 Optimization configuration

Before evaluating posterior inference performance, a small configuration study was conducted to select the optimization algorithm and regularization strength used throughout the remaining experiments. This step was necessary because the prior construction cost function is non-convex and, in preliminary experiments without τ regularization, the optimizer consistently drove τ toward arbitrarily large values, producing near-deterministic priors. Introducing an upper bound on τ prevented numerical divergence but did not resolve the underlying issue, as the optimizer frequently converged to the imposed upper bound. The variance-ratio regularization term was therefore introduced to discourage excessive prior concentration while retaining the ability to learn informative priors from the feasible region.

Powell and Nelder–Mead optimizers were compared under alternative optimization configurations. The selected configuration was evaluated on feasible-region compatibility, expected constraint violation, prior concentration, and cost function value. Table 4-7 summarizes the results.

Table 4-7: Comparison of optimization configurations for prior construction. Best-performing configuration highlighted in bold.

Optimizer	Penalty	w_τ	Prior mass ↑	Violation ↓	τ_{mean}	Cost function ↓
Powell	**variance_ratio_log**	**0.02**	**0.90–0.94**	**0.002–0.004**	high (∼70–110)	**−0.70 to −0.57**
Powell	variance_ratio_log	1.0	0.20–0.46	0.05–0.17	low (∼20)	0.26–0.84
Nelder–Mead	variance_ratio_log	0.02	0.44–0.57	0.02–0.05	very high (∼100–300)	0.03–0.10

Powell combined with the log variance-ratio penalty ($w_\tau = 0.02$) provides the best trade-off between high prior mass within the feasible region and low constraint violation. Strong

regularization ($w_\tau = 1.0$) produces overly diffuse priors, while Nelder–Mead exhibits higher variance and less consistent performance across restarts.

The cost function weights fixed throughout all simulation experiments are $w_v = 5.0$, $w_m = 1.0$, $w_d = 0.25$, and $w_\tau = 0.02$, reflecting the relative importance of constraint satisfaction, feasible-region coverage, centering, and concentration control. For each seed and regime, 10 optimization restarts are run and the solution with the lowest cost is selected:

$$\theta^* = \arg \min_{k=1,\dots,10} \mathcal{L}(\theta^{(k)}).$$

To ensure that $\mu_j \in (0,1)$ and $\tau_j \in [\tau_{\min}, \tau_{\max}]$ throughout optimization, the optimizer operates on an unconstrained vector $\boldsymbol{\eta} \in \mathbb{R}^{2J}$ mapped to $(\boldsymbol{\mu}, \boldsymbol{\tau})$ via the logistic transform:

$$\mu_j = \mu_\varepsilon + (1 - 2\mu_\varepsilon)\,\sigma(\eta_{2j}), \tag{4-8}$$

$$\tau_j = \tau_{\min} + (\tau_{\max} - \tau_{\min})\,\sigma(\eta_{2j+1}), \tag{4-9}$$

for $j = 0, \dots, J-1$, where $\sigma(x) = (1 + e^{-x})^{-1}$ and $\mu_\varepsilon > 0$ is a small constant keeping μ_j away from the boundary of $(0,1)$. This reparameterization ensures that all candidate solutions satisfy the prescribed bounds on μ_j and τ_j throughout optimization, while allowing the use of an unconstrained derivative-free optimizer. All computational settings are defined in Appendix B-2-1, Table B-1.

4-3-2 CPT posterior accuracy

This section evaluates posterior CPT accuracy under the optimized prior relative to the uniform baseline, across all eight target sizes and both constraint modes. Results are averaged over 15 independent random seeds.

Illustrative example: $N = 50$, strict constraints

To provide intuition into how the optimized prior influences inference, Figures 4-6 and 4-7 show the prior and posterior distributions and the posterior means with Wilson intervals for a representative seed at $N = 50$ under strict constraints.

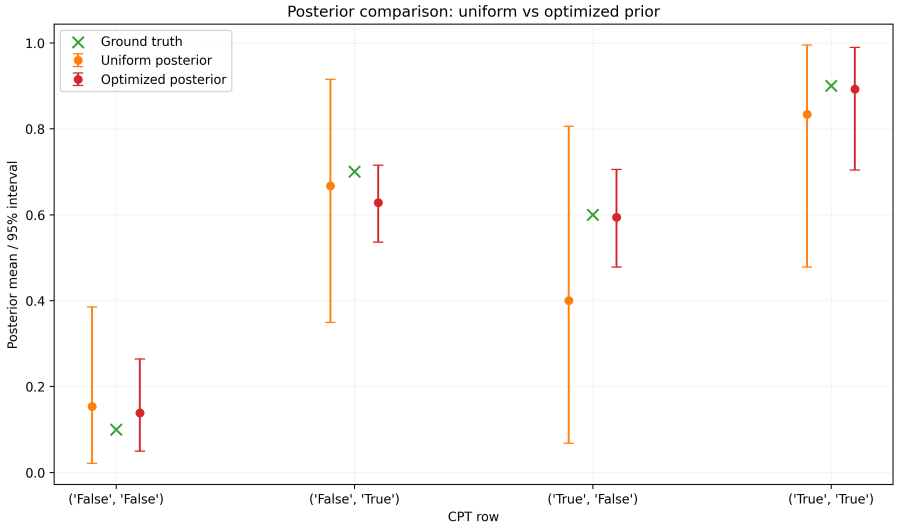


Figure 4-7: Posterior means and 95% credible intervals per CPT row for the representative seed ($N = 50$, strict constraints, seed 12). Ground-truth parameters are shown as dashed lines. The optimized prior reduces both posterior error and credible interval width across all four rows.

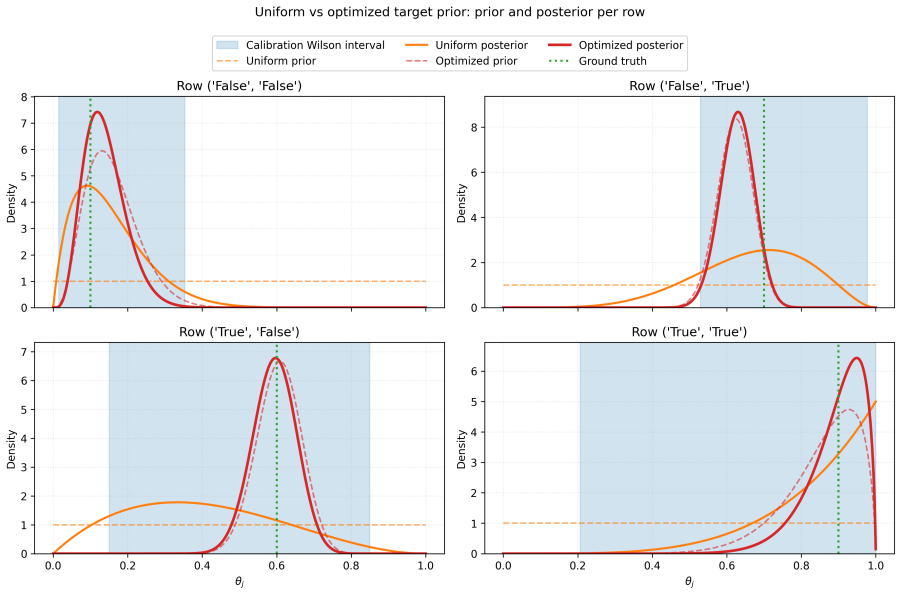


Figure 4-6: Prior and posterior Beta distributions per CPT row for the representative seed ($N = 50$, strict constraints, seed 12). The shaded region indicates the target-side Wilson interval. The optimized prior concentrates probability mass within the constraint-consistent feasible region, shifting the posterior toward the ground-truth parameter (dashed vertical line) relative to the uniform prior.

Table 4-8 reports the corresponding numerical results per row.

Table 4-8: Per-row inference results for the representative seed at $N = 50$, strict constraints (seed 12). $[L_j, U_j]$ is the Wilson calibration interval; $[a_j, b_j]$ is the marginal projection of $\mathcal{F}_{i,\text{mode}}$ after cross-row constraint intersection. τ denotes the concentration of the optimized Beta prior.

Row	θ^*	Calibration intervals		Uniform posterior			Optimized posterior				
		$[L_j, U_j]$	$[a_j, b_j]$	Mean	95% CI	Err	μ_{prior}	τ	Mean	95% CI	Err
(F, F)	0.10	[0.015, 0.354]	[0.015, 0.354]	0.154	[0.021, 0.385]	0.054	0.159	26.6	0.139	[0.049, 0.264]	0.039
(F, T)	0.70	[0.529, 0.978]	[0.529, 0.978]	0.667	[0.349, 0.915]	0.033	0.622	103.7	0.628	[0.536, 0.715]	0.072
(T, F)	0.60	[0.150, 0.850]	[0.379, 0.850]	0.400	[0.068, 0.806]	0.200	0.606	67.0	0.594	[0.478, 0.705]	0.006
(T, T)	0.90	[0.207, 1.000]	[0.529, 1.000]	0.833	[0.478, 0.995]	0.067	0.855	11.8	0.892	[0.704, 0.989]	0.008
Mean		–	–	–	–	0.088	–	52.3	–	–	0.031

The row-level results illustrate both the strengths and limitations of the framework. For the (F, F) row, the optimizer places the prior mean at 0.159, substantially closer to the true value ($\theta^* = 0.10$) than the uniform prior mean of 0.5 (Beta$(1, 1)$). Consequently, the posterior mean shifts from 0.154 to 0.139 and the absolute error decreases from 0.054 to 0.039. A similar effect is observed for the (T, T) row, where the optimized prior mean of 0.855 yields a posterior estimate of 0.892 and reduces the error from 0.067 to 0.008.

The (T, F) row demonstrates the additional benefit of cross-row constraints. While the Wilson interval alone spans $[0.150, 0.850]$, the feasible-region projection tightens this range to $[0.379, 0.850]$ through interaction with the monotonicity constraints. The optimized prior is therefore centered near the true value, producing a posterior mean of 0.594 and reducing the error from 0.200 to 0.006.

The (F, T) row demonstrates a limitation of the optimization approach. The optimizer assigns a high concentration parameter ($\tau = 103.7$), producing a narrow prior centered at 0.622. Although this prior remains feasible, its strong concentration limits the influence of the observed data and pulls the posterior toward the prior mean. As a result, the posterior estimate (0.628) remains farther from the true value of 0.70 than the corresponding uniform-posterior estimate (0.667).

This behavior is consistent with the coverage results reported above. When the prior mean is slightly displaced from the true parameter, a large τ reduces the influence of the observed data and can increase estimation error despite producing a narrower posterior. This high-τ behavior is systematic across seeds rather than specific to this example, as shown in the aggregate concentration values of Table B-2.

Aggregate results

Table 4-9 summarizes the MAE improvement and posterior credible-interval coverage of the optimized prior relative to the uniform baseline across all target sizes and both constraint modes, averaged over 15 seeds. Full results including RMSE, credible-interval width reduction, and KL divergence are reported in Appendix B-3.

Table 4-9: Summary of optimized prior performance relative to the uniform baseline across all target sizes and both constraint modes, averaged over 15 seeds. MAE improvement is the per-seed percentage $(\mathrm{MAE}_{\mathrm{uni}} - \mathrm{MAE}_{\mathrm{opt}})/\mathrm{MAE}_{\mathrm{uni}} \times 100\%$, averaged over seeds. Coverage is the proportion of CPT rows for which the true θ^* falls within the 95% posterior credible interval, reported for both the uniform (uni) and optimized (opt) priors.

	Strict ($\varepsilon = 0.15$)			Mild ($\varepsilon = 0.30$)		
N	MAE impr. (%)	Cov. (uni)	Cov. (opt)	MAE impr. (%)	Cov. (uni)	Cov. (opt)
10	47.5	1.000	0.667	55.5	1.000	0.833
20	36.6	0.983	0.633	41.1	0.983	0.750
30	32.3	0.950	0.700	35.2	0.950	0.817
50	32.2	0.933	0.733	35.4	0.933	0.800
75	33.8	0.983	0.800	32.2	0.983	0.833
100	14.4	0.967	0.733	20.1	0.967	0.833
150	24.7	0.917	0.717	34.5	0.917	0.833
200	19.2	0.917	0.800	23.4	0.917	0.817

The optimized prior consistently outperforms the uniform baseline across all target sizes under both constraint modes. Under strict constraints, MAE improvements range from 14.4% at $N = 100$ to 47.5% at $N = 10$, while under mild constraints improvements range from 20.1% to 55.5% (Table B-4 in Appendix B-3). Similar trends are observed for RMSE, credible interval width reduction, and Bernoulli KL divergence.

The largest improvements are observed at small target sample sizes. At $N = 10$, KL divergence is reduced by approximately 70% under strict constraints and more than 75% under mild constraints, indicating that the optimized prior not only improves point estimation accuracy but also produces posterior distributions that are substantially closer to the true parameter values. Consistent with Bayesian updating, the relative influence of the prior decreases as more target data becomes available. However, the performance gains do not decrease monotonically with N, suggesting that the optimized prior continues to exploit the qualitative constraints beyond simple low-sample regularization.

Across all target sizes, the mild constraint mode produces larger improvements than the strict mode. This follows from the geometry of the feasible region: the additional C_7 gap constraint under strict mode reduces the space available to the optimizer, forcing the prior mean closer to the boundary of $\mathcal{F}_{i,\mathrm{strict}}$ and limiting its ability to identify a well-centered solution.

Although posterior accuracy improves, calibration deteriorates. Under strict constraints, optimized-prior coverage ranges from 0.633 to 0.800, compared with 0.917–1.000 for the uniform baseline. Mild constraints achieve slightly higher coverage (0.750–0.833), but remain below the nominal 95% level. This reflects the tendency of the optimizer to construct high-τ priors that

concentrate probability mass near the projected center of the feasible region, improving point accuracy at the expense of uncertainty calibration.

The concentration values underlying this calibration loss are reported per row in Appendix B-3, Table B-2. Under strict constraints the optimizer assigns τ_{FT} and τ_{TF} values ranging from roughly 62 at $N = 10$ to over 120 at $N = 200$, far exceeding the effective sample size of the calibration split, which is what suppresses credible-interval coverage on those rows. The (T, T) row retains the lowest concentration throughout ($\tau_{TT} \approx 9$–33), because the (T, T) conditioning event occurs rarely and therefore little calibration data constrains that row. Strict constraints systematically inflate τ_{FT} and τ_{TF} relative to mild constraints, consistent with the tighter feasible region forcing the prior into a narrower admissible band.

4-3-3 Comparison with MPL-TC

In this stage, the optimized prior is compared against the three MPL-TC baselines described in Section 4-1-5: mild, moderate, and conflict source quality levels. The comparison is conservative by design, MPL-TC receives a source dataset of the same size as the target calibration split, no cross-row constraints are applied to its prior construction, and the source datasets are drawn independently from the families used in the metric validation experiment. Any observed advantage of the optimized prior therefore reflects the benefit of feasible-region-based construction alone, not a favorable experimental setup.

A critical practical distinction underlies this comparison. The MPL-TC baselines require access to a source dataset from a related domain. In operational HMF deployments, such a dataset may not exist, may represent a different context with unknown mismatch, or may introduce negative transfer if incorrectly selected. The optimized prior requires none of this: it is constructed entirely from the target calibration subset and qualitative monotonicity constraints that an analyst can specify from domain knowledge alone.

Figure 4-8 shows the MAE vs N_{target} trajectory for all methods under both constraint modes, and Table 4-10 reports the numerical values under strict constraints. Results under mild constraints are given in Appendix B-3.

Table 4-10: CPT posterior MAE under strict constraints across all target sizes, averaged over 15 seeds. Improvement percentages are computed relative to the optimized prior; negative values indicate the comparison method outperforms the optimized prior.

N	Uniform	MPL-TC mild	MPL-TC mod	MPL-TC conflict	Optimized	vs Uni (%)	vs mild (%)	vs mod (%)	vs conflict (%)
10	0.207	0.217	0.240	0.289	0.107	48.3	50.6	55.5	63.0
20	0.174	0.189	0.194	0.235	0.105	39.7	44.5	46.0	55.3
30	0.150	0.125	0.147	0.220	0.097	35.4	22.8	34.4	56.0
50	0.127	0.098	0.109	0.185	0.079	38.2	20.0	28.0	57.5
75	0.099	0.072	0.092	0.165	0.061	38.7	15.1	33.7	63.2
100	0.095	0.074	0.093	0.176	0.070	26.9	5.5	25.1	60.6
150	0.088	0.053	0.078	0.185	0.056	35.7	−6.6	28.1	69.6
200	0.063	0.049	0.060	0.158	0.047	25.5	4.5	21.9	70.2

Four findings stand out in Table 4-10 and Figure 4-8.

Strong low-N advantage. The optimized prior achieves the lowest mean MAE among all methods throughout $N \leq 100$ and remains competitive at larger sample sizes. The largest advantages over both the uniform baseline and the MPL-TC baselines occur in the low-data regime. At very small target sample sizes ($N \leq 20$), all three MPL-TC sources perform worse than the uniform baseline. Even at $N = 30$, only the mild source begins to provide a benefit, while the moderate and conflict sources remain comparable to or worse than the uniform prior. In contrast, the optimized prior achieves an MAE of 0.107 at $N = 10$, less than half the error

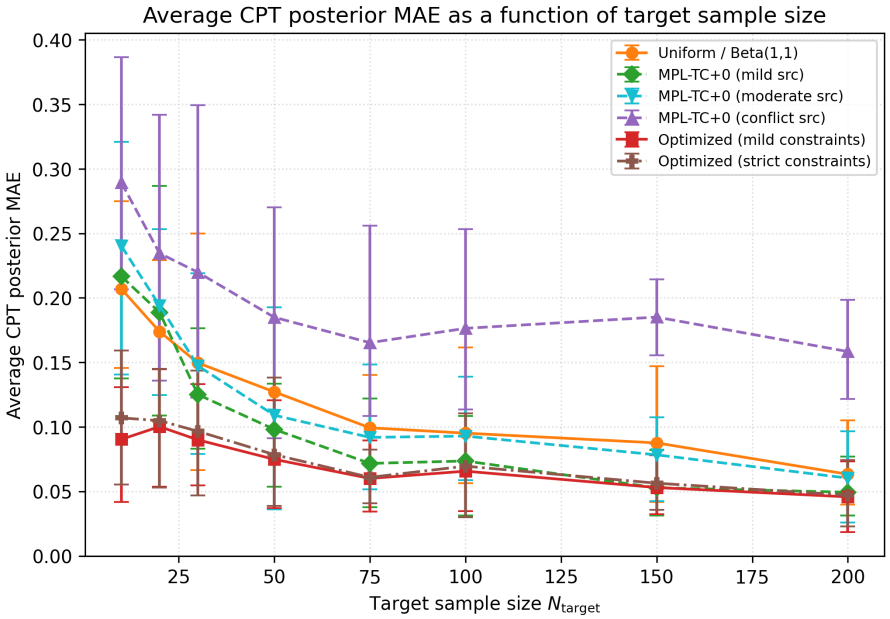


Figure 4-8: Posterior MAE as a function of target sample size for all methods, faceted by constraint mode. Shaded bands indicate ± 1 standard deviation across 15 seeds. The optimized prior (red) consistently achieves lower MAE than the uniform baseline (orange) at all N. The MPL-TC conflict source (dashed purple) remains the worst-performing method across the full N range.

of the conflict source (0.289). These results indicate that when both source and target datasets are extremely small, transfer learning may be less reliable than uninformed Bayesian learning. The optimized prior, by contrast, maintains substantially lower error throughout this regime.

MPL-TC mild converges to the optimized prior at large N. From $N \geq 100$ the advantage over MPL-TC mild is no longer statistically significant (all $p > 0.1$), and both methods achieve essentially equivalent MAE at large N. At $N = 150$ under strict constraints, MPL-TC mild achieves a marginally lower MAE (0.053 vs. 0.056), though the difference is not significant ($p = 0.76$). This convergence reflects an expected property: when the source domain is closely related to the target, transfer learning produces a prior that is itself nearly consistent with the feasible region, and both methods converge as the likelihood begins to dominate. This behavior is expected from Bayesian learning and serves as a useful sanity check for the proposed framework. The optimized prior achieves this convergence without requiring access to any source domain data.

Persistent advantage over the conflict source. The MPL-TC conflict source remains 55–70% worse than the optimized prior across all N values, with the gap widening rather than narrowing as N increases. This is an important finding: within the target-sample-size range considered here, increasing the amount of target data is insufficient to overcome the bias introduced by a severely misaligned source domain. The optimized prior is unaffected by source quality because it uses no source domain data at all.

KL divergence reveals source bias. Figure 4-9 shows the Bernoulli KL divergence from the true CPT parameters to the posterior mean estimate as a function of N. MPL-TC mild exhibits higher KL than the uniform prior at small to moderate N, despite achieving lower MAE. This apparent paradox reflects source bias: the mild source displaces the prior mean systematically, increasing distributional divergence from the truth even when it improves point estimation. The optimized prior achieves both lower mean MAE and lower mean KL throughout $N \leq 100$, indicating that the improvement extends beyond reductions in absolute estimation error and also yields posterior mean estimates that remain closer to the true probability distribution.

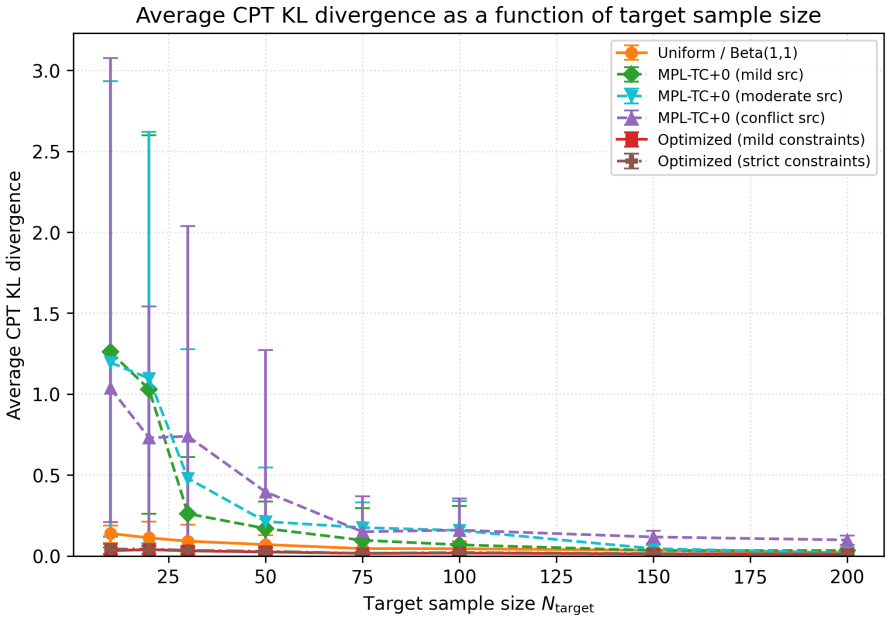


Figure 4-9: Bernoulli KL divergence from the true CPT parameters to the posterior mean estimate, as a function of target sample size, faceted by constraint mode. The MPL-TC mild source (green) exhibits higher KL than the uniform prior at small N despite achieving lower MAE, reflecting systematic source bias. The optimized prior (solid red) achieves the lowest KL throughout $N \leq 100$, indicating distributional accuracy. Shaded bands show the 10th–90th percentile range across 15 seeds rather than ± 1 standard deviation, since KL divergence is non-negative and right-skewed.

Paired Wilcoxon signed-rank tests (one-sided, $\alpha = 0.05$, $n = 15$ seeds per group) confirm that the optimized prior significantly outperforms the uniform baseline and the conflict sources across most target sizes and maintains statistically significant advantages over the mild and moderate MPL-TC sources primarily in the low-data regime. Full significance results are reported in Appendix B-3-3.

Compared with the uniform prior, reductions in CPT posterior MAE are statistically significant across nearly all target sizes under mild constraints and across all target sizes except $N = 200$ under strict constraints. This indicates that the observed mean-performance advantage is not merely a consequence of sampling variability across seeds.

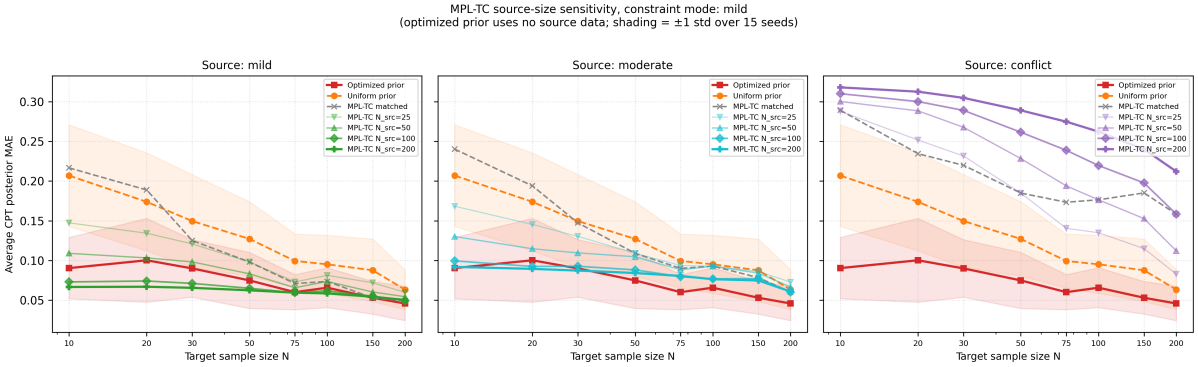
Taken together, a prior built from target calibration data and cross-row monotonicity constraints reduces CPT posterior MAE by 14–48% under strict constraints and 20–56% under mild constraints relative to the uniform baseline, and achieves lower mean MAE than the moderate and conflict MPL-TC baselines across the full target-sample-size range. While the optimized prior maintains lower mean MAE than MPL-TC mild through $N = 100$, the magnitude of this advantage becomes sufficiently small that statistical significance is retained only through $N = 30$. In contrast, the advantage in KL divergence remains statistically significant through $N = 100$ under both constraint modes, indicating that the optimized prior continues to produce posterior distributions that are more closely aligned with the true parameter values even after differences in point-estimation accuracy become difficult to distinguish statistically.

Source dataset size sensitivity for MPL-TC

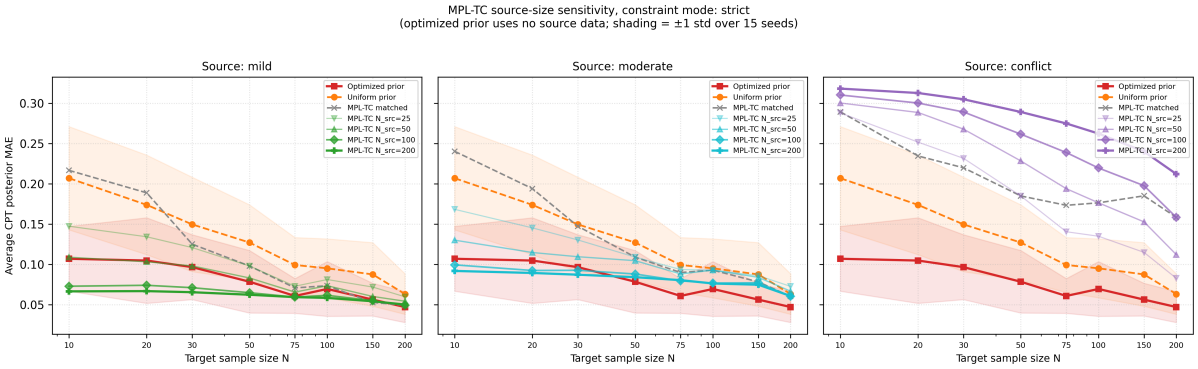
The results above use source datasets of the same size as the target calibration split ($N_{\text{src}} = N_{\text{target}}/2$), ensuring a fair comparison under an equal data availability assumption. In the original MPL-TC method, however, source domains contain 200–400 samples, substantially

more than the target, because the motivation for transfer is precisely that the source has richer data. To assess whether the optimized prior remains competitive when MPL-TC has access to a larger source, N_{src} is varied independently across 25, 50, 100, and 200 samples, while the target data size and the optimized prior construction remain unchanged.

Figure 4-10 shows CPT MAE as a function of N_{target} for each source quality and source size, under both constraint modes.



(a) Mild constraint mode.



(b) Strict mode.

Figure 4-10: MPL-TC source-size sensitivity. Each panel shows CPT posterior MAE vs target sample size for one source quality level (mild, moderate, conflict), with separate lines for each source dataset size N_{src}. The optimized prior (solid red) and uniform baseline (dashed orange) are shown as references; neither uses any source data. Shaded bands indicate ±1 standard deviation across 15 seeds. The optimized prior uses no source data at any point.

Three findings emerge. First, with a well-matched mild source and $N_{\text{src}} = 200$, MPL-TC does eventually fall below the optimized prior at larger N, with the crossover occurring at approximately $N_{\text{target}} = 75$–100 for mild constraints and $N_{\text{target}} = 150$–200 for strict constraints. At $N_{\text{src}} = 100$ the crossover already occurs at $N_{\text{target}} \approx 50$, confirming that the threshold for source transfer to outperform the feasible-region-based prior scales with both source size and constraint stringency.

Second, for the moderate source, larger N_{src} provides a benefit only at $N_{\text{target}} \leq 30$: even with $N_{\text{src}} = 200$, the moderate MPL-TC prior is outperformed by the optimized prior from $N_{\text{target}} \geq 50$ onward, with the gap reaching 30–40% at $N_{\text{target}} \geq 75$. A directionally plausible but imprecise source does not provide enough signal to overcome a well-calibrated optimized prior, except at very small target sample sizes.

Third, and most strikingly, larger source datasets make the conflict source *worse*: MPL-TC conflict with $N_{\text{src}} = 200$ reaches MAE values 197–363% above the optimized prior, compared to 124–249% in the matched-size comparison. A more confident wrong prior is more damaging

than a vague wrong prior. The optimized prior is entirely unaffected by this dynamic because it never consults any source domain data.

These results confirm that the primary advantage of the proposed approach is not merely that it matches MPL-TC under equal data availability, but that it provides reliable improvement regardless of whether a source domain exists, how large it is, or how well it matches the target. Transfer learning only becomes consistently competitive when a sufficiently large and well-aligned source dataset is available; in the present experiments this required source sample sizes on the order of 100 observations or more.

The following section examines whether these CPT-level improvements propagate to downstream network inference queries.

4-4 Network-level Inference Evaluation

The preceding section evaluates how well the proposed optimized prior recovers the true CPT parameters of the indicator node I. This section answers a different question: do improvements in CPT accuracy propagate to better hypothesis inference when the learned BBN is used to answer operational queries? This distinction matters because CPT error and query error have a non-linear relationship mediated by network structure. A small parameter improvement on a high-influence row can produce a large query improvement, while an equally large improvement on a rarely-activated row may have negligible downstream effect.

4-4-1 Experimental setup

The network evaluation uses the same experimental design as Section 4-3: 15 seeds, 8 target sizes, and 2 constraint modes. For each seed and regime, the posterior CPT means for $\mathbb{P}(I \mid H_1, H_2)$ are used to parameterize the indicator node, while all other nodes are fixed at their true parameter values. This isolates the contribution of the learned indicator node CPT parameters to downstream inference, ensuring that observed differences in query posteriors across methods are attributable to prior construction choices rather than variation in other network parameters.

Five inference queries are evaluated, covering both upward propagation from the observation node to the hypothesis nodes, and direct evidence at the indicator level:

- $\mathbb{P}(H_1 = T \mid S = T)$: hypothesis 1 posterior given a positive observation signal.
- $\mathbb{P}(H_2 = T \mid S = T)$: hypothesis 2 posterior given a positive observation signal.
- $\mathbb{P}(I = \text{Present} \mid S = T)$: indicator posterior given a positive signal.
- $\mathbb{P}(H_1 = T \mid I = \text{Present})$: hypothesis 1 posterior given indicator activation.
- $\mathbb{P}(H_2 = T \mid I = \text{Present})$: hypothesis 2 posterior given indicator activation.

For each query and each method, the absolute error relative to the true BBN query posterior is computed. Network MAE is the mean of these five absolute errors. A decision threshold of 0.5 is used to classify each query posterior as a positive or negative hypothesis assessment; decision error is the proportion of queries for which the learned BBN crosses this threshold in the wrong direction relative to the true BBN.

4-4-2 Query posterior accuracy

Figure 4-11 shows the network query MAE as a function of N_{target} for all methods under both constraint modes, and Table 4-11 reports the numerical values under strict constraints.

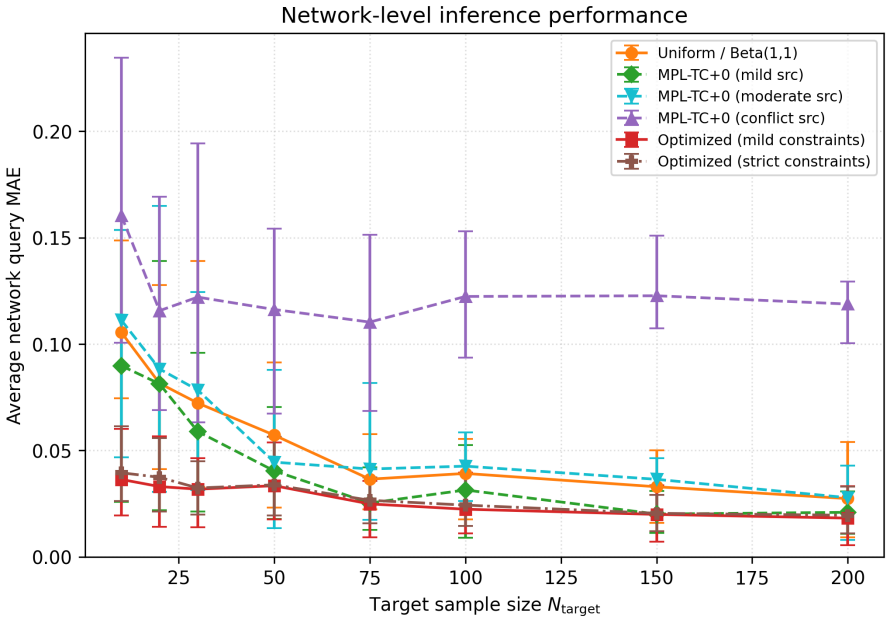


Figure 4-11: Average network query MAE as a function of target sample size, under both constraint modes. The optimized prior is shown for both constraint modes. The uniform baseline (dashed orange) and three MPL-TC sources (mild, moderate, conflict) are also shown. Shaded bands indicate ± 1 standard deviation across 15 seeds.

Table 4-11: Network query MAE under strict constraints across all target sizes, averaged over 15 seeds. Improvement percentages are computed relative to the optimized prior; negative values indicate the comparison method outperforms the optimized prior.

N	Uniform	MPL-TC mild	MPL-TC mod	MPL-TC conflict	Optimized	vs Uni (%)	vs mild (%)	vs mod (%)	vs conflict (%)
10	0.106	0.090	0.111	0.160	0.040	62.6	56.0	64.5	75.4
20	0.082	0.081	0.088	0.116	0.037	54.2	54.0	57.7	67.7
30	0.072	0.059	0.078	0.122	0.032	55.3	45.1	58.7	73.5
50	0.057	0.040	0.044	0.116	0.034	40.8	16.2	23.9	70.9
75	0.036	0.025	0.041	0.110	0.027	27.2	−5.1	35.7	75.9
100	0.039	0.031	0.043	0.122	0.024	38.0	22.7	43.0	80.2
150	0.033	0.020	0.036	0.123	0.020	37.9	−1.7	43.9	83.3
200	0.027	0.021	0.028	0.119	0.020	28.6	6.6	29.7	83.6

Figure 4-12 complements the MAE results by reporting Bernoulli KL divergence from the true network query posteriors to the estimated posteriors. The two metrics reveal the same ordering across methods but differ in emphasis. The MAE measures average absolute displacement in probability units and is bounded; the KL divergence is unbounded and penalizes confident wrong estimates disproportionately, amplifying cases where a biased prior places probability mass far from the truth near the distributional extremes.

This asymmetry explains why the MPL-TC conflict source appears far more catastrophic in the KL plot than in the MAE plot: at small N the conflict prior concentrates mass in a direction inconsistent with the true query posteriors, producing KL values an order of magnitude above the other methods (> 0.5 at $N = 10$) despite a MAE advantage of 75%. The optimized and uniform posteriors, by contrast, show comparable KL trajectories at small N, indicating that the MAE advantage of the optimized prior at small N reflects improved point location rather

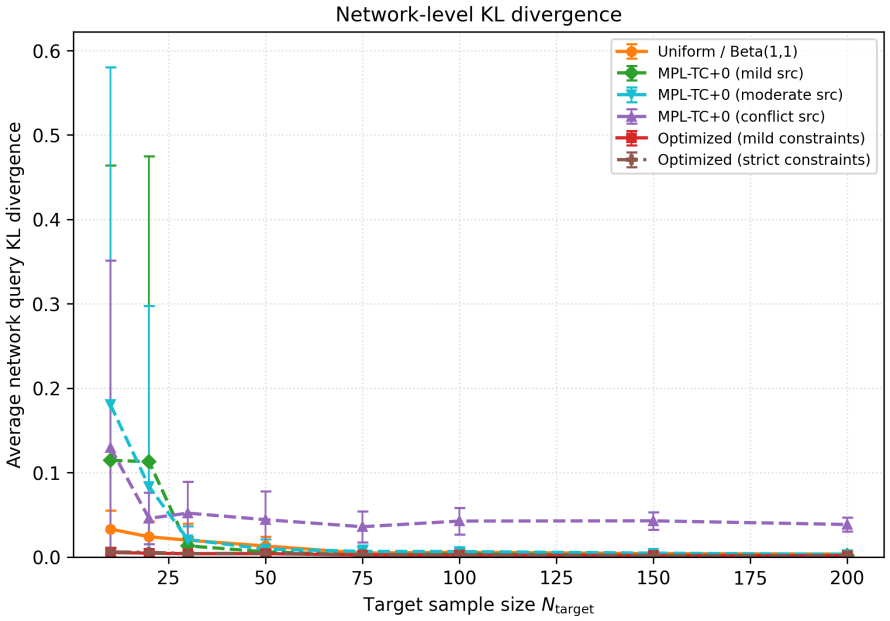


Figure 4-12: Bernoulli KL divergence from the true network query posteriors to the estimated posteriors, as a function of target sample size. MPL-TC sources are aggregated over constraint modes; the optimized prior is shown separately for mild (solid red) and strict (dash-dot dark red) constraints. Error bars indicate ± 1 standard deviation across 15 seeds.

than a fundamentally different distributional shape. At $N \geq 50$ the optimized prior achieves the lowest KL of all methods, consistent with the CPT-level KL results in Section 4-3-3.

The network results mirror the CPT results in two important respects but also reveal a key difference. The optimized prior consistently achieves the lowest query MAE across almost all N and both constraint modes, with improvements over the uniform baseline ranging from 27–63% under strict constraints. The conflict source remains the worst-performing method and its disadvantage grows with N: by $N = 200$ the optimized prior is 84% better than MPL-TC conflict, compared to 75% at $N = 10$. This widening gap is consistent with the conflict source becoming increasingly concentrated around incorrect parameter values as source size increases, making the resulting bias more difficult to overcome through Bayesian updating.

The key difference from the CPT results is the relative magnitude of the advantage. Network MAE improvements are substantially larger in percentage terms than the corresponding CPT MAE improvements. At $N = 10$ the CPT MAE improvement is 48% but the network improvement is 63%. This amplification occurs because the network queries depend on the joint configuration of all four CPT rows simultaneously, so even modest per-row improvements compound when propagated through the belief propagation mechanism.

Paired Wilcoxon signed-rank tests confirm that the optimized prior significantly outperforms the uniform baseline on network query MAE across most (N, mode) combinations (exceptions: $N = 75$ and $N = 200$ under strict constraints, $p > 0.05$). The advantage over the conflict source is significant across the full N range (all $p < 0.001$). The advantage over MPL-TC mild loses significance from $N = 50$ onward, consistent with the CPT-level convergence observed in Section 4-3-3. Full significance test results are reported in Appendix B-3-3.

4-4-3 Per-query analysis

Figure 4-13 shows the mean absolute error per query, aggregated across seeds at $N = 50$, strict constraints and Table 4-12 reports the corresponding numerical values.

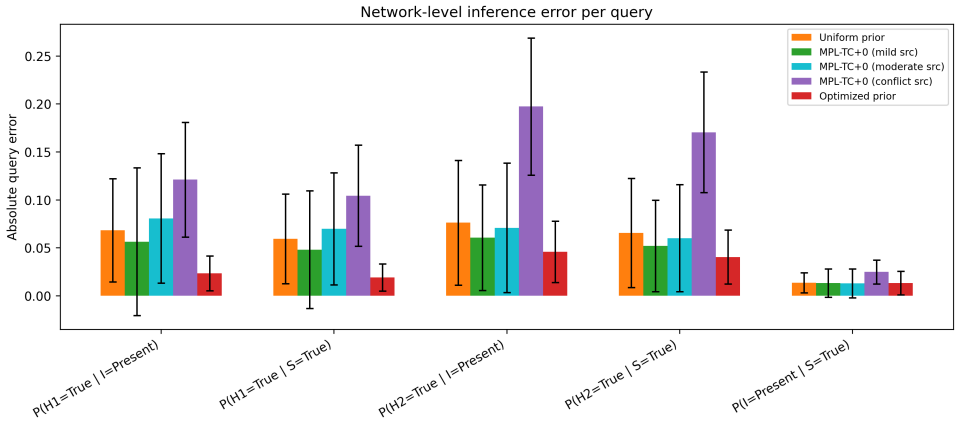


Figure 4-13: Mean absolute query error per inference query at $N = 50$, strict constraints, averaged over 15 seeds. Error bars indicate ± 1 standard deviation across seeds.

Table 4-12: Mean absolute query error per inference query at $N = 50$, strict constraints, averaged over 15 seeds.

Query	Uniform	MPL-TC mild	MPL-TC conflict	Optimized
$\mathbb{P}(H_1{=}T \mid S{=}T)$	0.065	0.047	0.102	0.024
$\mathbb{P}(H_2{=}T \mid S{=}T)$	0.061	0.042	0.159	0.045
$\mathbb{P}(I{=}\text{Present} \mid S{=}T)$	0.013	0.012	0.021	0.018
$\mathbb{P}(H_1{=}T \mid I{=}\text{Present})$	0.075	0.053	0.117	0.029
$\mathbb{P}(H_2{=}T \mid I{=}\text{Present})$	0.072	0.048	0.183	0.054
Mean	0.057	0.040	0.116	0.034

Two patterns emerge from Table 4-12. First, the $\mathbb{P}(I{=}\text{Present} \mid S{=}T)$ query shows negligible differences across methods (0.012–0.021). This query depends primarily on the observation node CPT $\mathbb{P}(S \mid I)$, which is fixed at its true value in this experiment, so the indicator CPT has limited influence. Second, the hypothesis queries show the largest absolute errors and the largest gains from the optimized prior. The H_2 queries show the largest absolute errors under MPL-TC conflict (0.159 and 0.183), suggesting that the conflict source prior is particularly poorly calibrated for the H_2 sub-network. These queries depend strongly on the indicator-node CPT and therefore benefit from the improvements in CPT estimation obtained through the feasible-region-based prior construction.

Query posterior uncertainty

Beyond point estimates of query error, the full distribution of query posteriors induced by CPT parameter uncertainty is characterized as described in Section 3-9-3. For each method, 500 CPT realizations are sampled from the Beta posteriors per seed, the BBN is instantiated with each realization, and all five queries are evaluated. Pooling across 15 seeds yields 7500 query posterior samples per method, from which a KDE is constructed. This represents the epistemic uncertainty over query answers that remains after observing $N = 50$ target samples under strict constraints.

Figure B-10 shows the resulting distributions for the uniform and optimized posteriors.

The KDE plots reveal that query posterior distributions overlap substantially between methods even when CPT-level MAE differences are large. This reflects the smoothing property of BBN inference: network queries are nonlinear functions of all four CPT rows simultaneously, weighted

by the marginal probabilities of parent configurations. With $\mathbb{P}(H_1{=}T) = 0.3$ and $\mathbb{P}(H_2{=}T) = 0.4$, the joint parent configuration (T, T) has only 12% prior probability weight, so even large differences in θ_{TT} between methods contribute limitedly to the marginal query posterior. The network structure therefore attenuates CPT-level differences at the query level, which explains why distributional separation is visible primarily through the mean location of each distribution relative to the true value rather than through differences in shape or width. The operationally relevant metric is therefore not the visual overlap of the distributions but the probability that each method places its query posterior on the correct side of the decision threshold.

Decision error analysis

Table 4-13 reports the mean decision error rate, the proportion of queries for which the posterior crosses the 0.5 threshold in the wrong direction, under strict constraints.

Table 4-13: Mean decision error rate (proportion of queries with incorrect binary classification at threshold 0.5) under strict constraints, averaged over 15 seeds. Bold indicates the lowest error at each N.

N	Uniform	MPL-TC mild	MPL-TC mod	MPL-TC conflict	Optimized
10	0.133	0.120	0.133	0.267	**0.080**
20	0.133	0.200	0.093	0.240	**0.067**
30	**0.080**	0.147	0.107	0.213	0.093
50	**0.027**	0.067	**0.027**	0.187	0.067
75	**0.013**	0.040	**0.013**	0.280	0.053
100	0.027	0.040	**0.000**	0.333	0.040
150	0.040	0.013	**0.000**	0.347	0.067
200	0.027	0.013	**0.000**	0.320	0.040

The decision error results present a more nuanced picture than the continuous MAE results. At very small N ($N \leq 20$), the optimized prior achieves the lowest decision error rate, reducing errors from 13% under the uniform prior to 6.7–8%. This is the operationally most important low-data regime: the constrained prior prevents the majority of threshold crossings that would cause an analyst to draw the wrong operational conclusion when data is severely limited.

From $N = 30$ onward, however, the decision error of the optimized prior does not decrease as rapidly as the uniform baseline and moderate MPL-TC source. At $N = 50$ the uniform prior and moderate source both achieve a decision error of 0.027, while the optimized prior remains at 0.067. This reversal reflects a direct consequence of the over-concentration effect: the optimizer places the prior mean at the projected center of the feasible region and assigns high τ values to concentrate probability mass there. As target data accumulates, the posterior mean is pulled toward this geometrically central point, which can sit near the 0.5 decision threshold for certain query configurations and produce threshold crossings that the data-driven uniform posterior avoids. The MAE metric is insensitive to this effect because errors on both sides of the threshold cancel in expectation, while the binary decision error detects it directly.

The MPL-TC moderate source achieves zero decision errors at $N \geq 100$. This indicates that, for this particular source-target mismatch scenario, the transferred prior remains sufficiently aligned with the decision boundary at larger sample sizes. The conflict source maintains a decision error rate of 0.187–0.347 throughout and remains the worst performer across the full N range.

The network-level results show that CPT-level improvements amplify when propagated through belief propagation, with network query MAE improvements of 27–63% exceeding the corre-

sponding CPT improvements. Decision error is reduced from 13% to 6.7–8% at $N \leq 20$, the operationally most important low-data regime. From $N \geq 30$ the uniform baseline achieves lower decision error than the optimized prior, reflecting the projected center concentration effect described above. The MPL-TC conflict source maintains a decision error rate of 0.187–0.347 throughout, while the optimized prior achieves 0.07–0.08 at small N.

4-5 Discussion

The results demonstrate that feasible-region-based prior construction from target calibration data and domain-informed constraints produces consistent CPT accuracy gains. Across all target sample sizes and constraint modes, the optimized prior reduces posterior CPT MAE by 14–56% relative to the uniform baseline. The largest improvements occur in the most data-scarce settings, where the prior exerts the strongest influence on the posterior. This suggests that the feasible region captures informative structural knowledge that cannot be recovered reliably from sparse observations alone, and that the proposed cost function successfully translates this information into improved prior distributions.

The mild constraint mode consistently produces larger improvements than the strict mode. The additional C_7 gap constraint reduces the size of the feasible region and restricts the set of admissible prior configurations. As a result, the optimization procedure has less freedom to locate a prior that simultaneously achieves high feasible-region prior mass, low expected violation, and a central location within the feasible region. While the strict constraints remain informative, the mild configuration appears to provide a more favorable balance between structural guidance and flexibility.

The optimized prior also compares favorably against transfer-learning baselines. It outperforms all three MPL-TC source priors despite requiring no source-domain data, with statistically significant MAE improvements through $N = 30$ and KL-divergence improvements through $N = 100$. When the source domain is well aligned with the target, the performance gap narrows as the likelihood increasingly dominates the posterior and the influence of the prior diminishes. In contrast, when the source domain is misaligned, additional source data do not resolve the problem. Instead, the source prior becomes increasingly concentrated around incorrect parameter values, making the resulting bias more difficult to overcome through Bayesian updating. Because the proposed framework relies exclusively on target-side information and domain constraints, it is unaffected by this source-mismatch effect. The source-size sensitivity analysis further showed that transfer learning only becomes consistently competitive when a sufficiently large and well-aligned source dataset is available.

Improvements at the parameter level propagate strongly to network-level inference. Across all target sample sizes, network query MAE improvements range from 27–66% across both modes, generally exceeding the corresponding gains observed at the CPT level. Since each network query depends jointly on all rows of the indicator-node CPT, small reductions in individual CPT estimation errors accumulate when translated into downstream inference, producing larger relative improvements in network-query MAE. The opposite effect is observed for the MPL-TC conflict source: parameter-level biases propagate through the network and produce a network MAE disadvantage of 68–84%, which becomes increasingly pronounced as the source prior concentration grows.

Several limitations qualify these findings. First, confidence interval coverage for the optimized prior remains below the nominal 0.95 level across all experimental settings, ranging from 0.633 to 0.833. This reflects a trade-off between calibration and point-estimation accuracy. The optimization procedure frequently selects large concentration parameters, producing priors that

place substantial probability mass near the projected center of the feasible region and thereby reducing posterior uncertainty.

Second, the same concentration effect influences decision-level performance. Although the optimized prior achieves the lowest decision error for $N \leq 20$, the uniform baseline performs better at larger sample sizes. The projected center provides a geometrically attractive reference point within the feasible region, but it is not guaranteed to coincide with the correct side of a downstream decision boundary. As highly concentrated priors pull posterior means toward this center, some query probabilities may cross the decision threshold in the wrong direction. While this effect has only a limited impact on MAE, which averages continuous prediction errors, it becomes visible when predictions are converted into binary decisions. Both the reduced confidence interval coverage and the decision-error reversal point to the same underlying phenomenon: an optimizer tendency toward excessive concentration. This issue is revisited in Chapter 5.

4-6 Summary

This chapter evaluated the proposed prior optimization framework through compatibility metric validation, CPT parameter learning, and downstream network inference. The method constructs a row-wise Beta prior by minimizing a composite cost function over feasible-region coverage, constraint violation, and distance to the marginal projected center of the feasible region, using target calibration data and qualitative monotonicity constraints without access to source-domain data.

These results confirm that the compatibility metrics consistently identify the most compatible and least compatible source families across target sizes, constraint modes, and random seeds, and that the violation decomposition correctly attributes global incompatibility to specific rows and cross-row ordering relationships.

At the CPT level, the optimized prior reduces posterior MAE by 14–48% under strict constraints and 20–56% under mild constraints relative to the uniform baseline, with the largest gains occurring at small N. Against the uniform baseline, CPT MAE improvements are statistically significant across nearly all target sizes under both constraint modes. Against MPL-TC transfer learning, the optimized prior outperforms the moderate and conflict sources across the full range and remains competitive with the mild source, with both methods converging at larger N, despite requiring no source data. The MAE advantage over the mild source is significant at small target sample sizes and the KL advantage persists over a larger range of N, after which both methods converge. The conflict source remains 55–70% worse in CPT MAE across the full N range, with the gap widening as N and source size increase.

At the network level, CPT-level improvements amplify into network query MAE reductions of 27–66% across both constraint modes over the uniform baseline, and the optimized prior achieves the lowest decision error at $N \leq 20$. From $N \geq 30$ onward the uniform prior achieves lower decision error, reflecting the over-concentration effect identified above. The MPL-TC conflict source maintains a decision error rate of 0.187–0.347 throughout, while the optimized prior achieves 0.07–0.08 at $N \leq 20$.

The principal limitation is the trade-off between point accuracy and calibration: high-τ priors improve MAE and KL but reduce CI coverage and, at the network level, can increase decision error at moderate and large N. Potential approaches for mitigating this effect are discussed in Chapter 5.

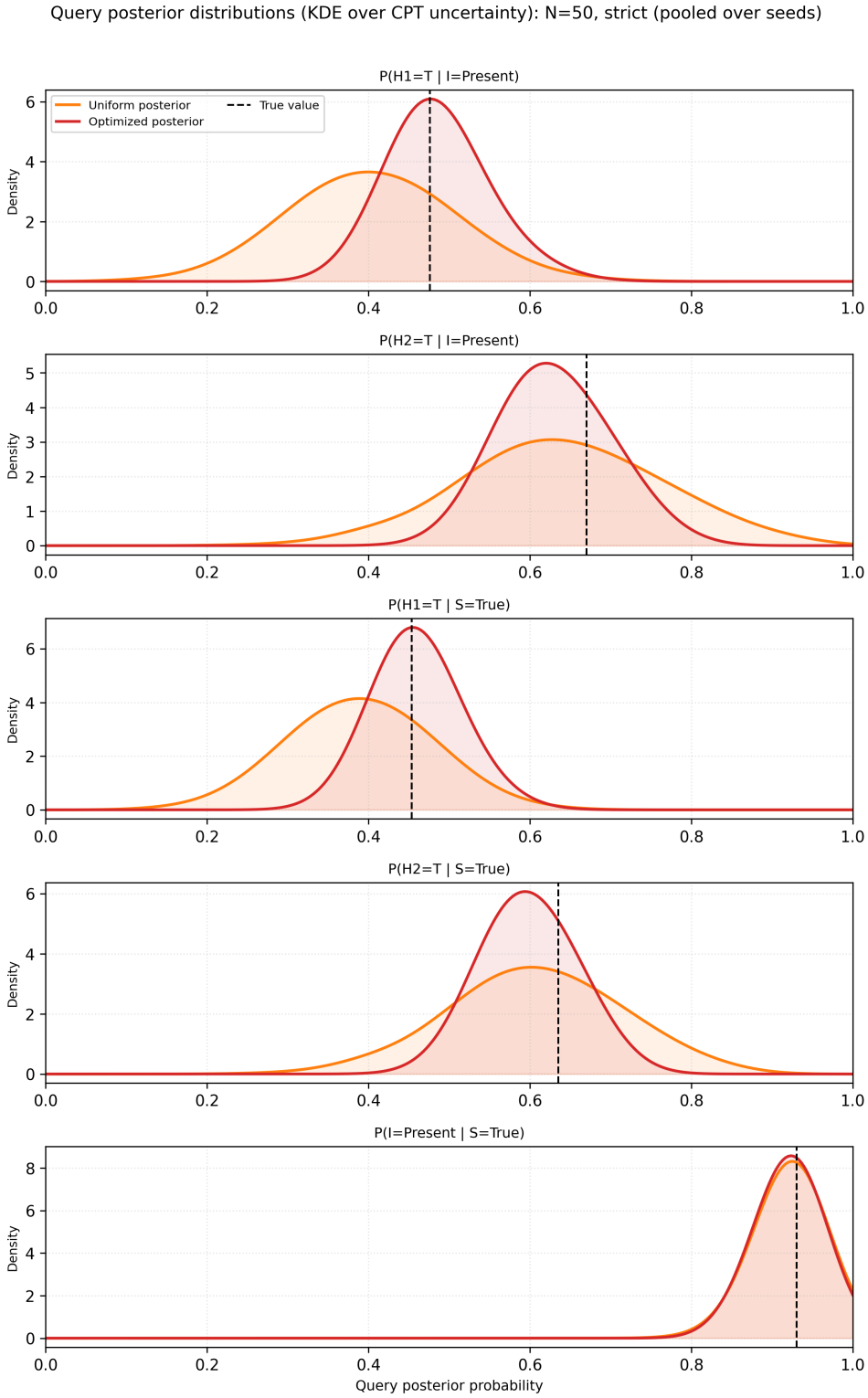


Figure 4-14: Query posterior distributions represented as KDE curves, pooled over 15 seeds at $N = 50$, strict constraints. Each panel shows one of the five inference queries. The dashed vertical line indicates the true query posterior under the ground-truth BBN. The optimized posterior (red) is more concentrated and its mean is consistently closer to the true value than the uniform posterior (orange), reflecting reduced parameter uncertainty and improved distributional placement. The full comparison including all MPL-TC sources is shown in Appendix B-4-2.

Chapter 5

Discussion and Conclusion

This thesis investigated whether it is possible to develop an approach for evaluating and refining prior knowledge in BBNs under data scarcity, using feasible parameter regions defined by target-side uncertainty and domain-informed constraints.

This chapter integrates the findings of the research and answers each research question in light of the experimental results reported in Chapter 4. It begins by repeating the main objective of the thesis and summarizing the main results.

Sections 5-2 through 5-4-3 discuss the results in depth, connect them to the related work reviewed in Chapter 2, and acknowledge the principal limitations per research question. Section 5-5 concludes the thesis with direct answers to the research questions and an outline of concrete directions for future work.

5-1 Overall Thesis Conclusion

Research Objective

"The objective of this thesis is to determine whether target-side uncertainty and qualitative domain knowledge can be transformed into informative Bayesian priors that improve parameter learning and inference in BBNs under data scarcity, without relying on source-domain transfer."

The objective of this thesis was to develop and evaluate a framework for Bayesian prior construction in BBNs under data scarcity, using target-side uncertainty and domain-informed constraints to construct informative priors without relying on source-domain data.

The results indicate that the proposed framework successfully addresses the research objective within the scope of the simulated experiments. A feasible-region-based prior optimization framework was developed that combines data-driven uncertainty intervals with qualitative cross-row constraints to define a structured set of admissible parameter configurations. Compatibility metrics were introduced to evaluate how well candidate priors align with this feasible region and were subsequently incorporated into a composite cost function for prior construction.

The resulting framework was evaluated through controlled simulation experiments and compared with both uniform priors and transfer-learning baselines. The optimized priors consistently improved posterior CPT estimation accuracy in data-scarce settings and yielded corresponding improvements in network-level inference. In particular, the framework achieved

statistically significant reductions in posterior MAE at small target sample sizes while requiring no source-domain information. These gains propagated through the network, resulting in substantial reductions in query error and decision error in the most data-scarce regimes.

The results demonstrate that target-side uncertainty and qualitative domain knowledge can be transformed into informative Bayesian priors through feasible-region-based optimization. This provides a practical alternative to source-based transfer learning in settings where suitable source domains are unavailable, difficult to validate, or potentially misleading.

With this, the overall goal of the thesis was met. The process showed several insights and limitations, which are discussed in the following sections for each of the three main research questions.

5-2 RQ1: Feasible-region construction

Research Question 1

"How can target-side uncertainty and qualitative knowledge be combined into a coherent feasible parameter region for BBN CPTs?"

5-2-1 Result

The feasible parameter region $\mathcal{F}_i$ is constructed by intersecting Wilson score confidence intervals with a set of linear cross-row monotonicity constraints encoding domain knowledge (Sections 3-2-1 through 3-2-3). At the target sizes evaluated (N = 10–200), the resulting region is non-empty across all seeds and experimental conditions, confirming the regularity condition stated in Section 3-2-3. The region contracts monotonically with the target data size N in the row-wise direction and becomes more selectively shaped as constraints tighten, separating source families that are indistinguishable under the credal box alone (Section 4-2-3, Figure 4-2).

5-2-2 Insights

The use of Wilson score intervals to define data-driven bounds on CPT parameters follows the interval estimation approach reviewed in Section 2-2-4. Unlike the imprecise-probability literature, which typically treats the credal set as the primary object of inference, this thesis treats the credal box $\mathcal{B}_i$ as a construction device for the prior rather than as a posterior summary. This represents a different use of credal-box representations than is typically found in the BBN parameter-learning literature.

The integration of cross-row constraints follows the constrained parameter-learning tradition, which imposes ordering constraints as linear inequalities on the admissible parameter space. The key extension here is that these constraints are applied jointly with the data-driven interval to produce a single geometric object $\mathcal{F}_i$ that serves as the reference for prior evaluation and construction. The resulting polytope structure is exploited in the centering regularization, which is not possible when constraints are incorporated only as soft penalties during posterior updating.

5-2-3 Limitations

The framework relies on the regularity condition that $\mathcal{F}_i$ is non-empty, which holds in practice at small N but could fail under very large samples combined with strict constraints if the

empirical proportion lies near a constraint boundary. No fallback procedure is defined for this case. The binary parameterization limits applicability to Bernoulli CPT rows; the generalization to multinomial rows via Dirichlet-Multinomial conjugacy is left for future work. Wilson score interval estimation is well-defined for any row count, but its informativeness degrades as the count falls: rows with few observations yield very wide intervals, and a parent configuration that is entirely unobserved ($N_{ij} = 0$) yields the maximal interval $[0, 1]$ (Assumption 5). In networks with many parent configurations and severely limited data, many rows may fall into this low-count regime, so the data-driven component of the feasible region provides little constraint and the region is shaped predominantly by the cross-row constraints. In this low-count regime the framework therefore depends most heavily on the correctness of those constraints: when data are too sparse to constrain a row, an incorrectly specified constraint cannot be corrected by the likelihood, an interaction examined further in the constraint-correctness limitation of Section 5-4-3.

5-3 RQ2: Compatibility quantification and decomposition

Research Question 2

"How can the compatibility of candidate prior knowledge with the feasible region be quantified and interpreted?"

5-3-1 Result

The two compatibility metrics, prior mass $\hat{m}(\mathbb{P}_s)$ and expected constraint violation $\mathbb{E}_{\theta\sim\mathbb{P}_s}[V(\theta)]$, generally reproduce the expected ordering of source prior families across target sizes and constraint modes, while consistently separating well-aligned sources from strongly conflicting ones (Section 4-2-3, Table 4-6). Discriminative power increases substantially as cross-row constraints tighten: without cross-row constraints, the feasible region reduces to the credal box and intermediate families become nearly indistinguishable at small N; adding cross-row constraints separates them clearly. The ranking remains stable across seeds (Figure B-3), confirming metric robustness. Joint evaluation detects structural incompatibilities, such as a source prior satisfying each Wilson interval individually while violating a cross-row monotonicity constraint, that row-wise assessment misses entirely (Section 4-2-2).

The violation decomposition (Section 3-5, illustrated in Section 4-2-5) localizes global incompatibility to specific CPT rows and constraint types. For a conflict source, the decomposition reveals that constraints C_1, C_2, C_4 and C_7 (involving θ_{FF} and θ_{TT}) contribute the largest violation shares, providing an analyst with a targeted diagnosis without requiring inspection of the full source prior distribution.

5-3-2 Insights

The mass-based compatibility metric is conceptually related to the prior-data conflict diagnostic, which identifies prior-data conflict as a regime in which the prior assigns most of its mass to parameter values for which the observed data is surprising. The present framework adapts this perspective to a constrained parameter space: infeasibility is driven not only by data conflict but also by structural constraint violation. Decomposing this broader notion of incompatibility into its two sources, interval violation and cross-row constraint violation, is a novel contribution that appears to be less commonly addressed in the BBN parameter-learning literature reviewed in Chapter 2.

The finding that discriminative power is amplified by cross-row constraints has a direct implication for the design of BBN models for decision support: analysts should invest in eliciting qualitative ordering constraints even when precise prior probabilities are unavailable, because these constraints substantially increase the informativeness of the compatibility metrics. This corroborates the argument that qualitative constraints can substitute for precise numerical elicitation in data-scarce settings.

5-3-3 Limitations

The prior mass metric treats all feasible configurations as equally admissible, which means that a prior with high mass concentrated near the boundary of $\mathcal{F}_i$ scores identically to one concentrated at the interior. The centering distance term in the cost function addresses this limitation in the construction mode, but the evaluation metric itself does not distinguish the two cases. While the metrics reliably distinguish highly compatible and strongly conflicting priors, discrimination between intermediate mismatch levels is weaker than at the extremes (Table 4-6), which limits the practical utility of the metrics as a fine-grained selection tool when multiple source priors have similar compatibility scores. An extended set of metrics, such as the conditional expected violation or a distributional divergence from the posterior predictive, may improve discriminative power in this regime.

5-4 RQ3: Prior construction and downstream inference

Research Question 3

"How can compatibility diagnostics be used to construct or refine priors under data scarcity, and what is the impact on downstream BBN inference?"

5-4-1 Result

The optimized prior reduces CPT posterior MAE by 14–56% across both modes relative to the uniform baseline and achieves lower mean MAE than the moderate and conflict MPL-TC sources across the full range of target sample sizes, while remaining competitive with the mild source. The advantage over MPL-TC mild is statistically significant through $N = 30$ on MAE and through $N = 100$ on KL divergence, after which both methods converge as the likelihood dominates. The advantage over the conflict source is statistically significant across the full N range on all three metrics (CPT MAE, KL, and network-query MAE; all $p < 0.001$, Figures B-8 and B-9). Larger source datasets amplify rather than mitigate the harm of a conflict source. As source dataset size increases, the transferred prior becomes increasingly concentrated around incorrect parameter values, causing CPT posterior MAE to increase rather than decrease (Figure 4-10). The performance gap relative to the optimized prior therefore widens with source size.

At the network level, CPT-level improvements amplify when translated into downstream inference. Network-query MAE improvements of 27–66% across both modes generally exceed the corresponding CPT-level improvements because each query depends jointly on the four rows of the indicator-node CPT (Section 4-4-2, Table 4-11). Small reductions in individual CPT estimation errors therefore accumulate through the network and produce larger relative improvements in query accuracy. The optimized prior achieves the lowest decision error at $N \leq 20$ ($\leq 8\%$, compared to 13% for the uniform prior), confirming operationally relevant gains in the most data-scarce regime (Table 4-13).

5-4-2 Insights

A prior built entirely from target-side information and qualitative constraints can achieve performance comparable to, and in many cases better than, the evaluated transfer-learning baselines in the studied data-scarce scenarios, without requiring any external data. This directly addresses the limitation of source-dependent transfer learning identified in Section 2-4. The MPL-TC framework assumes access to a related source domain with richer data; when the source is misaligned, this assumption introduces negative transfer. The proposed framework avoids this failure mode by construction because it relies exclusively on target-side information and qualitative constraints.

The convergence of the optimized prior with MPL-TC mild at large N reflects an expected and desirable property: as the likelihood dominates, all reasonable priors produce similar posteriors, and the relative advantage of any structured prior shrinks. This convergence suggests that any bias or variance induced by the optimized prior diminishes as the likelihood becomes increasingly dominant.

An additional observation is that the optimized prior remains highly competitive with the mild MPL-TC source even at the largest target sample sizes considered. Under both constraint modes, the difference in posterior MAE becomes negligible by $N = 150 - 200$, with both methods achieving nearly identical performance. This suggests that a substantial portion of the benefit traditionally attributed to transfer learning can, in some settings, be recovered from target-side uncertainty and qualitative constraints alone. Consequently, the proposed framework may serve as a practical substitute when a suitable source domain is unavailable or difficult to validate.

More broadly, this result indicates that the framework is most beneficial in the regime for which it was designed: severe data scarcity. When only limited target observations are available, the feasible-region constraints provide information that would otherwise be unavailable to the learning procedure. As target sample size increases, the influence of prior regularization naturally decreases and the quality of uncertainty calibration becomes increasingly important relative to point-estimation accuracy.

Improvements at the parameter level propagate strongly to network-level inference. Since each query depends jointly on multiple CPT parameters, small reductions in individual parameter errors accumulate through the belief propagation process, producing larger relative gains in network-query accuracy than at the CPT level. The per-query analysis further shows that the magnitude of these gains varies across inference tasks, indicating that some network queries are more sensitive to CPT parameter-estimation errors than others.

5-4-3 Limitations

Three systematic limitations qualify these results.

Over-concentration and calibration loss. The cost function rewards feasible-region mass and proximity to the projected center of $\mathcal{F}_i$, which creates an incentive to assign high concentration τ to rows where the prior mean can be placed close to the center. The resulting high-τ priors achieve low MAE and KL divergence but reduce posterior credible interval coverage to 0.633–0.833 across conditions, well below the nominal 0.95 level. The concentration regularization term $\mathcal{R}_\tau$ mitigates but does not eliminate this tendency. A practitioner who requires calibrated uncertainty quantification should treat the optimized prior as a starting point and apply an additional calibration step, such as coverage-targeted tuning of w_τ. This same tendency toward excessive concentration also appears to contribute to the decision-error reversal discussed below.

Decision-error reversal at moderate N. The optimized prior achieves the lowest decision error only at $N \leq 20$. From $N = 30$ onward, the uniform baseline achieves lower decision error (Table 4-13). This result illustrates that improvements in parameter accuracy do not necessarily translate into improvements in threshold-based decision making. The reversal appears to be driven by the same over-concentration effect observed in the calibration analysis. The cost function favors prior means located near the interior of the feasible region while assigning relatively large concentration parameters. Although this improves average parameter accuracy, the resulting posterior means can be pulled toward values that lie close to a decision boundary. Small shifts in the posterior can then move a query probability across the 0.5 threshold, producing an incorrect binary decision despite a lower continuous prediction error. This limitation is most relevant in operational HMF deployments where binary hypothesis assessments are required alongside continuous posterior probabilities.

Constraint correctness assumption. The framework assumes the imposed cross-row constraints are satisfied by the true CPT. If a constraint is incorrect, i.e. slightly violated by the true parameter, the projected center and violation term are computed against a feasible region that excludes the true value, biasing prior construction toward an incorrect region; because the prior mean is drawn toward this biased center, the resulting estimate inherits the bias, and under sparse data the likelihood may be unable to correct it, as discussed for the low-count regime in Section 5-2. The robustness of the approach to such misspecification is not evaluated in this thesis. This question has been addressed directly for a soft-penalty constrained learner by deliberately mislabelling a random subset of elicited monotonic influences and measuring the resulting degradation [56]; an analogous sensitivity analysis for the feasible-region construction, which relies on a hard region intersection for its centering reference, is identified as future work.

5-5 Conclusion and future work

This thesis developed and evaluated a feasible-region-based Bayesian prior optimization framework for CPT parameter learning in BBNs under data scarcity. The main contribution is a data-efficient procedure that represents target-side uncertainty and qualitative domain knowledge as a unified feasible parameter region and formulates prior construction as an optimization problem over that region. The framework provides a structured alternative to uniform priors that does not require access to a source domain, does not assume precise numerical prior elicitation from experts, and is robust to source-domain mismatch by construction. The framework therefore provides a practical alternative in settings where suitable source domains are unavailable, difficult to validate, or potentially misleading.

Answers to the research questions

RQ1. A feasible parameter region is constructed by intersecting Wilson score confidence intervals derived from the target calibration data with linear cross-row monotonicity constraints encoding domain knowledge. The resulting bounded convex polytope $\mathcal{F}_i$ is non-empty across all experimental conditions, contracts monotonically as data accumulates, and becomes more selective as constraints tighten, confirming that it provides an informative and well-behaved representation of admissible CPT configurations.

RQ2. Prior-region compatibility is quantified by prior mass and expected constraint violation, estimated via Monte Carlo sampling. These metrics generally reproduce the expected ordering of source families and consistently distinguish well-aligned priors from strongly conflicting ones. Joint feasible-region evaluation detects structural incompatibilities that row-wise assessment

misses, and the violation decomposition localizes global incompatibility to specific CPT rows and constraints, enabling targeted analyst intervention in intelligence analysis.

RQ3. A prior constructed entirely from target-side data and constraints reduces CPT posterior MAE by 14–56% over the uniform baseline, without requiring any source-domain data. It outperforms the moderate and conflict MPL-TC baselines across the full range of target sizes and remains competitive with the mild source, with the two converging at large N as the likelihood dominates. CPT-level improvements amplify to 27–66 % network query MAE reductions and reduce decision error from 13 % to below 8 % at $N \leq 20$. The over-concentration tendency of the optimizer reduces posterior calibration and introduces a decision-error reversal at moderate N, identifying adaptive concentration regularization and threshold-aware cost function design as the most important directions for future work.

5-5-1 Directions for future work

- The fixed weight w_τ in the cost function does not adjust to the growing influence of the likelihood as N increases. An adaptive scheme that increases w_τ with N could preserve the decision-error advantage of the optimized prior beyond $N = 20$ while maintaining MAE gains at small N.
- For HMF deployments where binary hypothesis assessments are operationally required, the cost function could incorporate a penalty on prior means placed near the 0.5 decision threshold for the relevant query configuration. This would directly target the decision-error failure mode identified at moderate N.
- The current framework isolates parameter uncertainty to a single indicator node. Extending to simultaneous estimation of all CPTs requires defining feasibility constraints per node and accounting for confounding between node-level uncertainties. A natural approach is to optimize row-wise priors per node independently and evaluate joint query performance.
- The framework is restricted to binary nodes under a Beta-Binomial model. The natural generalization replaces this with a Dirichlet-Multinomial formulation, where the feasible region is defined over the probability simplex. This extension would substantially broaden applicability to networks with non-binary discrete variables.
- The simulation study validates the framework under known ground truth. Deploying it on real-world HMF data would demonstrate practical utility under conditions where the true CPT is unknown and evaluation must rely on predictive performance and consistency with domain-informed constraints.
- The framework assumes that imposed constraints are satisfied by the true CPT. A robustness analysis evaluating performance when one or more constraints are slightly violated by the true parameters would characterize the sensitivity of the approach to constraint quality and provide practical guidance on constraint specification.
- Future work could investigate more principled mappings between interval-based uncertainty representations and prior concentration. In particular, the equal-weight averaging of the lower and upper Wilson interval widths adopted in this thesis could be replaced by asymmetric weighting schemes or posterior variance estimates, potentially yielding improved calibration of the optimized prior.
- The observed decision-error reversal may partly depend on the choice of a fixed 0.5 threshold. In operational intelligence settings, decision thresholds are often context-dependent

and reflect asymmetric costs associated with missed detections and false alarms. Future work could therefore investigate threshold-sensitive optimization and evaluation procedures.

Appendix A

Simulation-based experiment

A-1 Example derivation of the query $P(H_1 = \mathrm{T} \mid I = \mathrm{Present})$

The optimized model learns the conditional probability table

$$P(I \mid H_1, H_2),$$

where H_1 and H_2 are binary hypothesis nodes and I is a binary indicator node.

The four CPT parameters are defined as

$$\theta_{FF} = P(I = \mathrm{Present} \mid H_1 = \mathrm{F}, H_2 = \mathrm{F}),$$

$$\theta_{FT} = P(I = \mathrm{Present} \mid H_1 = \mathrm{F}, H_2 = \mathrm{T}),$$

$$\theta_{TF} = P(I = \mathrm{Present} \mid H_1 = \mathrm{T}, H_2 = \mathrm{F}),$$

$$\theta_{TT} = P(I = \mathrm{Present} \mid H_1 = \mathrm{T}, H_2 = \mathrm{T}).$$

To evaluate the effect of the learned CPT on downstream inference, network-level probabilistic queries are computed. One example is

$$P(H_1 = \mathrm{T} \mid I = \mathrm{Present}),$$

which can be derived using Bayes' rule.

Assuming that H_1 and H_2 are independent root nodes with prior probabilities

$$p_1 = P(H_1 = \mathrm{T}), \qquad p_2 = P(H_2 = \mathrm{T}),$$

Bayes' rule gives

$$P(H_1 = \mathrm{T} \mid I = \text{Present}) = \frac{P(I = \text{Present} \mid H_1 = \mathrm{T})P(H_1 = \mathrm{T})}{P(I = \text{Present})}.$$

Since the CPT depends on both H_1 and H_2, the term

$$P(I = \text{Present} \mid H_1 = \mathrm{T})$$

is obtained by marginalizing over H_2:

$$P(I = \text{Present} \mid H_1 = \mathrm{T}) = \sum_{h_2} P(I = \text{Present} \mid H_1 = \mathrm{T}, h_2)P(h_2).$$

Substituting the binary states of H_2 yields

$$P(I = \text{Present} \mid H_1 = \mathrm{T}) = (1 - p_2)\theta_{TF} + p_2\theta_{TT}.$$

The denominator is obtained by marginalizing over all states of H_1 and H_2:

$$P(I = \text{Present}) = \sum_{h_1}\sum_{h_2} P(I = \text{Present} \mid h_1, h_2)P(h_1)P(h_2).$$

Expanding the binary states gives

$$P(I = \text{Present}) = (1 - p_1)(1 - p_2)\theta_{FF} + (1 - p_1)p_2\theta_{FT} + p_1(1 - p_2)\theta_{TF} + p_1p_2\theta_{TT}.$$

Combining numerator and denominator results in

$$P(H_1 = \mathrm{T} \mid I = \text{Present}) = \frac{p_1\left[(1 - p_2)\theta_{TF} + p_2\theta_{TT}\right]}{(1 - p_1)(1 - p_2)\theta_{FF} + (1 - p_1)p_2\theta_{FT} + p_1(1 - p_2)\theta_{TF} + p_1p_2\theta_{TT}}.$$

In the experiments, the root-node priors are initialized symmetrically:

$$P(H_1 = \mathrm{T}) = P(H_2 = \mathrm{T}) = 0.5.$$

Under this assumption, the expression simplifies to

$$P(H_1 = \mathrm{T} \mid I = \text{Present}) = \frac{\theta_{TF} + \theta_{TT}}{\theta_{FF} + \theta_{FT} + \theta_{TF} + \theta_{TT}}.$$

This illustrates that the posterior belief in H_1 depends on the relative contribution of the CPT rows in which $H_1 = \mathrm{T}$ compared to the total probability mass assigned to the indicator state $I = \text{Present}$.

Appendix B

Supplementary Results for Simulation Experiments

This appendix contains supplementary analyses, diagnostic plots, and complete numerical results assisting the findings reported in Chapter 4. The material is included to improve reproducibility and provide additional insight into the behavior of the feasible region construction, prior optimization procedure, CPT parameter-learning results, and downstream network inference performance.

B-1 Feasible region construction and compatibility metrics

This section provides additional visualizations of the feasible region introduced in Chapter 3. Because the feasible region is defined in a four-dimensional parameter space, direct visualization is not possible. The projections shown below illustrate how the Wilson-score credal box is reduced by the monotonicity and closeness constraints.

The purpose of these figures is not to analyze a specific experimental setting, but to provide intuition regarding the geometry of the feasible region. In particular, they demonstrate that the imposed constraints introduce dependencies between CPT rows, transforming the independent Wilson intervals into a structured admissible parameter set.

B-1-1 Feasible region projections

B-1-2 Compatibility metric validation

The compatibility metrics are intended to distinguish between priors that are well aligned with the target feasible region and priors that are inconsistent with the available data and cross-row constraints. This section provides supplementary results validating that the metrics behave as expected across a range of source-target mismatch scenarios.

The results demonstrate that increasing source-target mismatch is associated with lower feasible-region mass and higher expected constraint violation, confirming that the proposed metrics capture meaningful differences in compatibility.

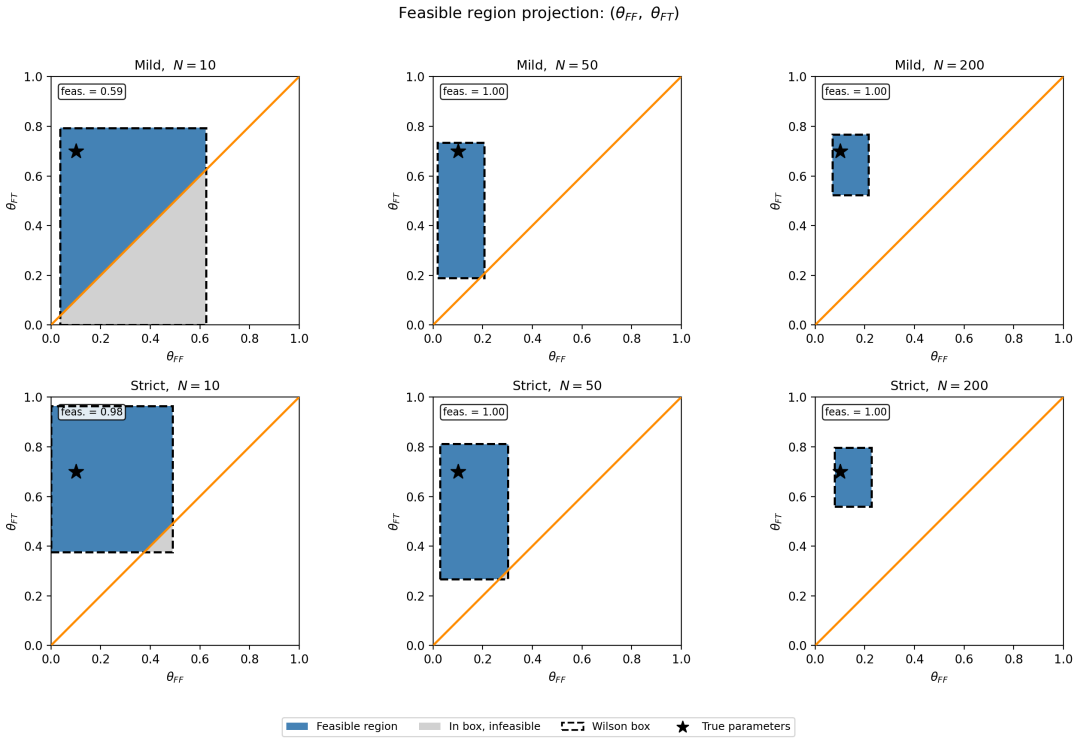


Figure B-1: 2D projection of the feasible region $\mathcal{F}$ onto the $(\theta_{FF}, \theta_{FT})$ plane for a representative experiment. The outer rectangle corresponds to the Wilson-score credal box $\mathcal{B}$, while the shaded region denotes the subset satisfying all cross-row constraints. The projection illustrates how qualitative monotonicity and closeness constraints reduce the admissible parameter space relative to the unconstrained uncertainty intervals.

Because the compatibility metrics are estimated from finite Monte Carlo samples, an important question is whether the resulting source ranking is stable across random seeds. The following heatmap evaluates the robustness of the ranking procedure.

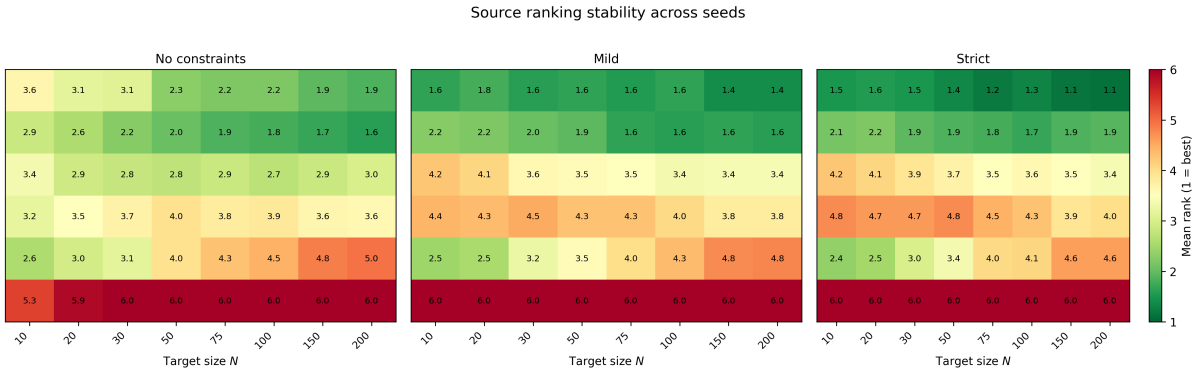


Figure B-3: Source family ranking stability across 20 random seeds, under strict constraints at $N = 100$. Each cell shows the frequency with which a source family achieves a given rank position by prior mass in the feasible region, across seeds. Darker cells indicate more consistent ranking. The equal source is top-ranked and the conflict source is bottom-ranked in the large majority of seeds, while intermediate families show greater variability.

While prior mass and expected violation provide interpretable diagnostics individually, they can be combined into a single compatibility score. Figure B-4 illustrates how the base cost function (Equation (3-29)), combining feasible-region mass, expected violation, and centering,

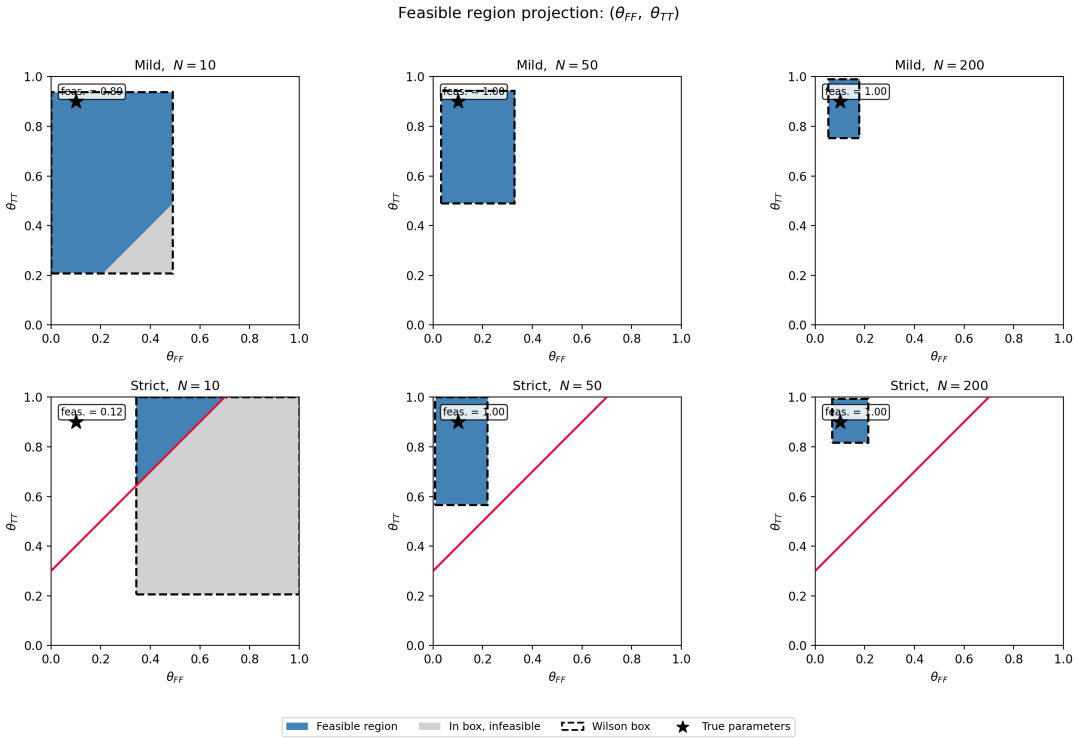


Figure B-2: 2D projection of the feasible region $\mathcal{F}$ onto the $(\theta_{FF}, \theta_{TT})$ plane for the same experiment. The feasible region forms a structured subset of the Wilson-score credal box due to the imposed cross-row constraints. This projection highlights the dependence introduced between parameters that would otherwise be represented independently through row-wise uncertainty intervals.

ranks representative source families. The τ-regularization term is omitted here, as it governs the concentration of a constructed prior and has no role in scoring fixed source priors.

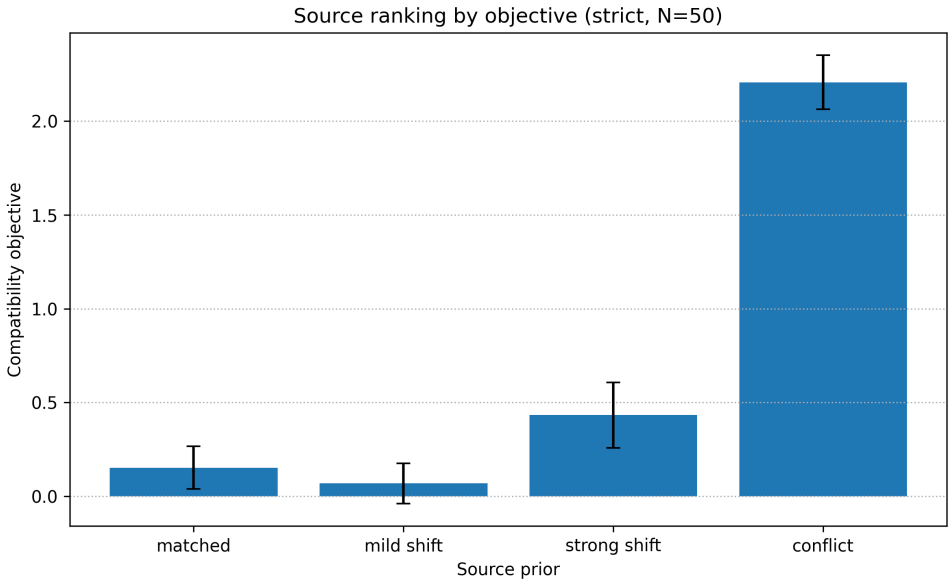


Figure B-4: Mean compatibility base cost function (mass + violation + distance) value for the four representative source priors under strict constraints at $N = 50$, averaged over 20 seeds. Lower values indicate greater compatibility with the feasible region. The matched and mild-shift priors achieve the lowest cost values, while the conflict prior is consistently assigned the highest penalty. Error bars denote one standard deviation across seeds. The ranking demonstrates that the base cost function successfully discriminates between increasing levels of prior-target mismatch.

B-1-3 Incompatibility localization

The previous analysis quantifies overall incompatibility but do not identify where the incompatibility originates. The figures in this section decompose the violation metric into individual constraints and CPT rows, providing insight into the structural source of prior-target mismatch.

Such diagnostic information is useful because two priors may exhibit similar overall compatibility scores while violating entirely different parts of the feasible region.

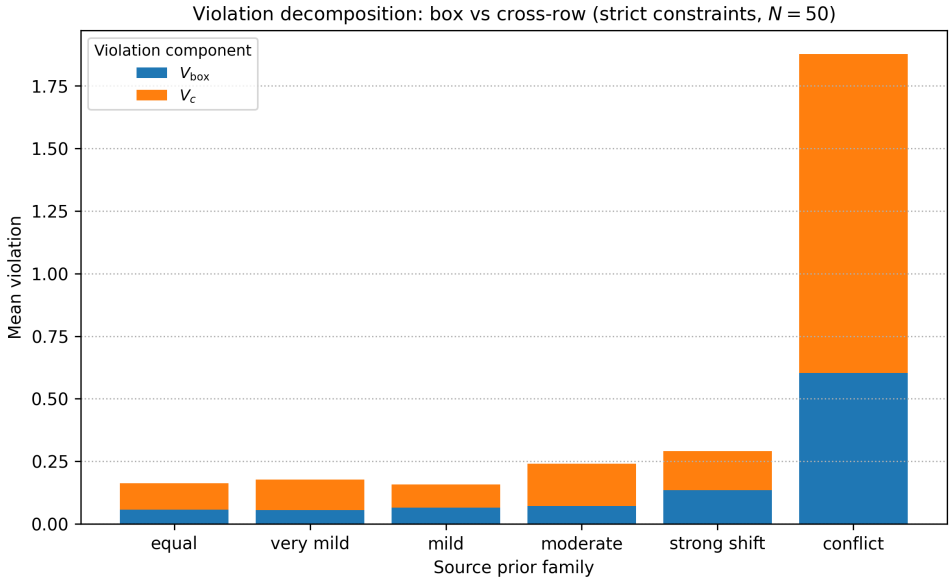


Figure B-5: Mean violation decomposition across six source prior families under strict constraints ($N = 50$). Each bar separates the total mean violation into box violation V_{box} (prior mean outside the Wilson calibration interval) and structural violation V_c (cross-row constraint violations). The equal source produces near-zero violation on both components. Conflict produces dominant structural violation from monotonicity reversal, while strong shift accumulates primarily box violation from location mismatch. Mild and moderate sources show mixed contributions, illustrating that equal total violation can arise from qualitatively different structural sources.

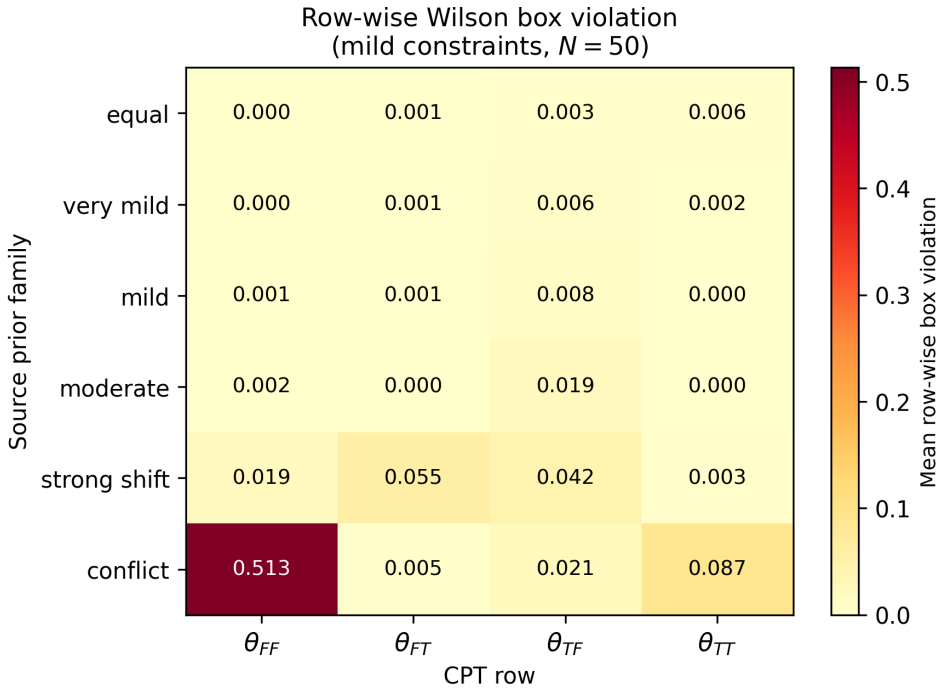


Figure B-6: Row-wise box violation heatmap for all source families at $N = 50$, mild constraints, averaged over 20 seeds. Each cell reports the mean violation of that CPT row against its Wilson interval bound. The conflict source concentrates violation almost entirely on θ_{FF} (0.513), indicating that its main row-wise mismatch is the inflated baseline probability assigned to the (F, F) parent configuration. The remaining rows show substantially smaller violations. The strong shift source produces a more moderate violation profile.

B-2 Prior optimization

B-2-1 Computational settings

All experiments were implemented in Python using `scipy.optimize.minimize`. Table B-1 summarizes the key settings. Powell's derivative-free algorithm was selected because the cost function contains Monte Carlo estimates and indicator-based feasibility terms, making analytical gradients unavailable.

Table B-1: Computational settings for the prior optimization experiments.

Setting	Value
Optimizer	
Algorithm	Powell (derivative-free, direction-set method)
Maximum iterations	300
Parameter tolerance x_{tol}	10^{-3}
Function tolerance f_{tol}	10^{-3}
Multi-start	
Number of restarts	10
Odd starts	Wilson interval midpoints ($\mu_j = \frac{L_j+U_j}{2}$, $\tau_{\text{init}} = 10$)
Even starts	Fully random; seed $= 4000 + \text{seed}_{\text{run}}$
Monte Carlo samples	
Cost function evaluation (per cost function call)	750 samples per call
Final evaluation (post-optimization)	10 000 samples
Query distribution plots	500 posterior draws per BBN
MPL-TC bootstrap replicates	100
Cost function weights	
Constraint violation penalty w_{viol}	5.0
Prior mass weight w_m	1.0 (enters as $-w_m \hat{m}$, maximized)
$\mathcal{F}_i$ projection center distance weight w_d	0.25
Concentration regularization w_τ	0.02
Penalty type	Variance-ratio log
Maximum τ	500
Data and seeds	
Calibration/inference split	50 % / 50 % (stratified by conditioning row)
Split seed	$2000 + \text{seed}_{\text{run}}$
Cost function MC seed	$3000 + \text{seed}_{\text{run}}$
Number of independent runs	15
Run seeds	$0, 1, \ldots, 14$
Target dataset subsets	Nested prefix slices of a common draw of $N_{\max} = 200$
Inference	
Posterior credible intervals	95 % equal-tailed Beta quantiles

The nested prefix design ensures that datasets at smaller N are strict subsets of the dataset at $N = 200$ within the same seed, eliminating sampling variability as a confound when comparing results across sample sizes. Cost function function evaluations are stochastic (Monte Carlo integration over the joint feasible region $\mathcal{F}$); fixing `seed_objective` to $3000 + \text{seed}_{\text{run}}$ ensures that each run sees the same Monte Carlo draws throughout its optimization trajectory.

These settings were selected based on preliminary experiments balancing optimization stability and computational cost. The multi-start strategy reduces sensitivity to local optima, while the Monte Carlo sample sizes provide sufficiently stable cost function estimates without excessive runtime.

The purpose of this section is primarily reproducibility; all experiments reported in Chapter 4 were conducted using the settings summarized in Table B-1.

B-2-2 Optimization behavior

The optimized priors are parameterized through Beta distributions whose concentration parameter $\tau = \alpha + \beta$ controls the strength of the prior relative to the likelihood. Examining the optimized τ values provides insight into how the optimizer balances feasible-region compatibility against prior uncertainty.

Table B-2 reports the resulting concentration parameters across target sample sizes and constraint modes.

Table B-2: Mean optimized concentration parameters $\tau_j = \alpha_j + \beta_j$ per CPT row, averaged over 15 seeds. Values in parentheses are standard deviations. Strict constraints systematically inflate τ_{FT} and τ_{TF} relative to mild constraints, while τ_{TT} remains lowest due to sparse calibration data in the (T, T) conditioning event.

N	Mode	τ_{FF}	τ_{FT}	τ_{TF}	τ_{TT}
10	mild	15.2 (13.5)	29.4 (7.5)	26.2 (3.9)	9.3 (2.8)
10	strict	15.2 (13.7)	62.2 (9.7)	57.3 (10.4)	10.4 (2.4)
30	mild	18.6 (7.5)	35.2 (12.0)	28.2 (8.9)	9.2 (2.5)
30	strict	18.3 (7.2)	71.8 (20.0)	63.4 (14.1)	10.2 (2.2)
50	mild	23.3 (9.4)	48.2 (22.7)	38.7 (16.3)	13.4 (9.3)
50	strict	20.0 (9.2)	82.5 (25.6)	74.4 (23.4)	13.7 (8.2)
100	mild	45.5 (9.2)	55.9 (23.2)	36.2 (9.4)	19.5 (7.6)
100	strict	47.4 (9.6)	101.7 (32.5)	76.2 (11.4)	18.4 (7.5)
200	mild	85.1 (13.2)	78.6 (22.6)	49.6 (7.3)	32.8 (12.2)
200	strict	84.8 (12.5)	121.3 (21.2)	87.8 (15.7)	33.2 (8.2)

Several trends are visible. First, concentration generally increases with target sample size, reflecting the shrinking Wilson intervals and the increasingly precise estimate of the feasible region. Second, the strict constraint mode consistently produces larger concentration parameters for the middle CPT rows (FT and TF), indicating that the additional constraint reduces feasible region flexibility and encourages more concentrated solutions. Finally, the (T, T) row remains the least concentrated across all settings, reflecting the limited amount of calibration data available for this conditioning configuration.

B-3 CPT parameter-learning results

B-3-1 Representative feasible-region-based optimization result

The main rapport evaluates the optimized priors quantitatively through posterior inference performance. The figure below provides a qualitative illustration of a representative optimization outcome.

The visualization shows the feasible region, the optimized prior mean, and the resulting concentration of probability mass for a single representative experiment. It illustrates how the optimization balances constraint satisfaction, feasible region coverage, and proximity to the projected center of the feasible region.

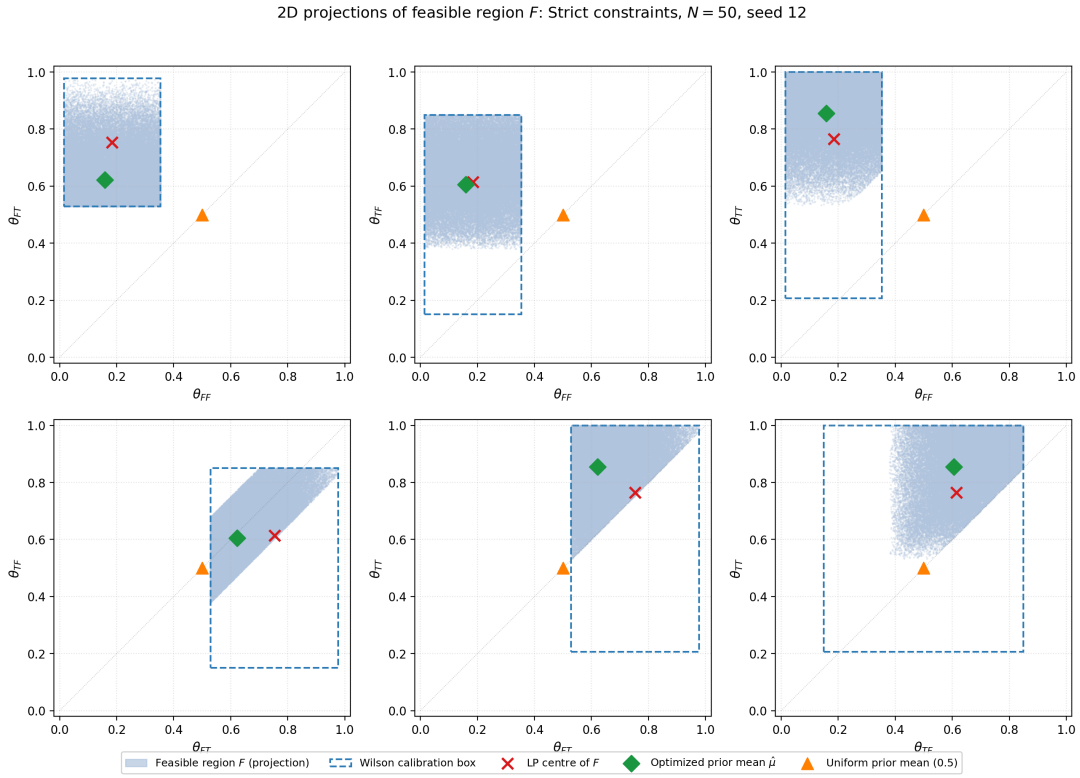


Figure B-7: Representative optimization result for a single experimental run ($N = 50$, strict constraints, seed 12). The shaded region denotes the feasible parameter set $\mathcal{F}$ obtained from the Wilson-score intervals and cross-row monotonicity constraints. The projected center of the feasible region serves as the reference point in the distance component of the cost function, while the optimized prior concentrates probability mass within the admissible region. The figure illustrates how the optimization balances feasible region coverage, constraint satisfaction, and proximity to the representative interior of $\mathcal{F}$.

B-3-2 Full CPT comparison tables

The main chapter focuses on relative performance improvements and statistical significance. For completeness, the full numerical results are reported below.

These tables provide the exact posterior inference metrics underlying the figures presented in Section 4-3-3.

Table B-3: CPT posterior MAE for mild and strict constraints across all target sizes, averaged over 15 seeds. Baseline columns are identical across constraint modes. The optimized prior achieves lower MAE under mild constraints throughout, as the less restrictive feasible region allows greater flexibility in prior placement.

Mode	N	Uniform	MPL-TC mild	MPL-TC mod	MPL-TC conflict	Optimized	vs Uni (%)	vs mild (%)	vs mod (%)	vs conflict (%)
Mild	10	0.207	0.217	0.240	0.289	0.090	56.3	58.3	62.4	68.8
	20	0.174	0.189	0.194	0.235	0.100	42.4	46.9	48.4	57.3
	30	0.150	0.125	0.147	0.220	0.090	39.9	28.1	38.9	59.1
	50	0.127	0.098	0.109	0.185	0.075	41.1	23.7	31.4	59.5
	75	0.099	0.072	0.092	0.165	0.060	39.5	16.2	34.6	63.7
	100	0.095	0.074	0.093	0.176	0.066	31.0	10.7	29.3	62.8
	150	0.088	0.053	0.078	0.185	0.053	39.5	−0.3	32.3	71.4
	200	0.063	0.049	0.060	0.158	0.046	27.7	7.3	24.1	71.1
Strict	10	0.207	0.217	0.240	0.289	0.107	48.3	50.6	55.5	63.0
	20	0.174	0.189	0.194	0.235	0.105	39.7	44.5	46.0	55.3
	30	0.150	0.125	0.147	0.220	0.097	35.4	22.8	34.4	56.0
	50	0.127	0.098	0.109	0.185	0.079	38.2	20.0	28.0	57.5
	75	0.099	0.072	0.092	0.165	0.061	38.7	15.1	33.7	63.2
	100	0.095	0.074	0.093	0.176	0.070	26.9	5.5	25.1	60.6
	150	0.088	0.053	0.078	0.185	0.056	35.7	−6.6	28.1	69.6
	200	0.063	0.049	0.060	0.158	0.047	25.5	4.5	21.9	70.2

Table B-4: CPT posterior inference performance across all target sizes and both constraint modes, averaged over 15 seeds. MAE and RMSE improvements are per-seed percentage improvements averaged over 15 seeds. CI reduction is the mean reduction in 95% posterior credible interval width. KL is the mean Bernoulli KL divergence from the true parameter to the posterior mean. Coverage is the proportion of CPT rows for which θ^* falls within the 95% posterior CI.

Mode	N	MAE_{uni}	MAE_{opt}	MAE impr. (%)	RMSE impr. (%)	CI reduction	KL_{uni}	KL_{opt}	Cov_{uni}	Cov_{opt}
Strict	10	0.207	0.107	47.5	46.0	0.500	0.139	0.042	1.000	0.667
	20	0.174	0.105	36.6	36.3	0.428	0.112	0.043	0.983	0.633
	30	0.150	0.097	32.3	34.7	0.374	0.091	0.034	0.950	0.700
	50	0.127	0.079	32.2	32.3	0.323	0.070	0.028	0.933	0.733
	75	0.099	0.061	33.8	32.8	0.278	0.045	0.016	0.983	0.800
	100	0.095	0.070	14.4	22.9	0.231	0.045	0.021	0.967	0.733
	150	0.088	0.056	24.7	27.6	0.203	0.034	0.014	0.917	0.717
	200	0.063	0.047	19.2	26.6	0.154	0.021	0.010	0.917	0.800
Mild	10	0.207	0.090	55.5	53.0	0.449	0.139	0.032	1.000	0.833
	20	0.174	0.100	41.1	40.2	0.377	0.112	0.039	0.983	0.750
	30	0.150	0.090	35.2	35.7	0.328	0.091	0.031	0.950	0.817
	50	0.127	0.075	35.4	36.9	0.288	0.070	0.025	0.933	0.800
	75	0.099	0.060	32.2	36.2	0.254	0.045	0.015	0.983	0.833
	100	0.095	0.066	20.1	30.8	0.201	0.045	0.017	0.967	0.833
	150	0.088	0.053	34.5	39.7	0.181	0.034	0.012	0.917	0.833
	200	0.063	0.046	23.4	29.2	0.135	0.021	0.011	0.917	0.817

B-3-3 Statistical Significance Tests

Paired one-sided Wilcoxon signed-rank tests are conducted per (N_{target}, *constraint mode*) group to assess whether observed performance differences are statistically reliable rather than due to sampling variability across seeds. The Wilcoxon test does not assume normally distributed differences, making it appropriate for small samples ($n = 15$ seeds) and right-skewed metrics such as KL divergence. The null hypothesis is that the optimized prior does not outperform the baseline; a significance level of $\alpha = 0.05$ is used throughout

Figures B-8 and B-9 report the full results of the paired one-sided Wilcoxon signed-rank tests described in Section 4-3-3. Each cell shows the p-value when the result is not significant ($p \geq 0.05$); green cells indicate that the optimized prior significantly outperforms the baseline ($p < 0.05$), red cells indicate no significant difference at that N.

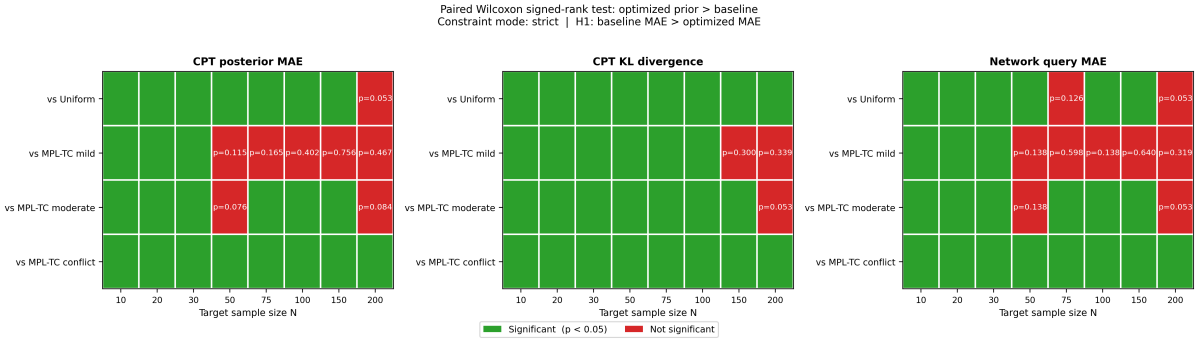


Figure B-8: Paired Wilcoxon signed-rank test results under strict constraints, for all comparisons (optimized prior vs uniform, MPL-TC mild, moderate, and conflict) across all target sample sizes N and three metrics: CPT posterior MAE (left), CPT KL divergence (centre), and network query MAE (right). $n = 15$ seeds per group. Green cells indicate a statistically significant advantage of the optimized prior over the baseline ($p < 0.05$); red cells show the p-value where significance is not reached. The optimized prior significantly outperforms all baselines at $N \leq 30$ on CPT MAE and network MAE. The KL divergence advantage over MPL-TC mild is maintained through $N = 100$, after which both methods converge. The conflict source comparison remains significant across the full N range on all three metrics.

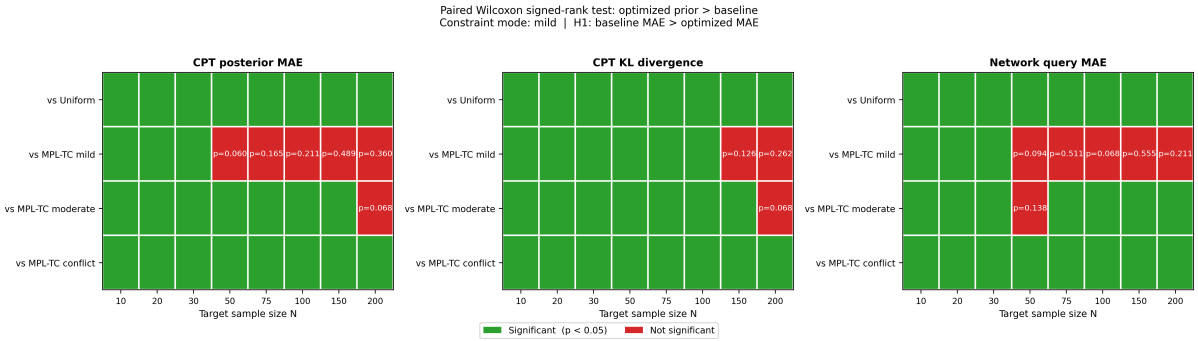


Figure B-9: Paired Wilcoxon signed-rank test results under mild constraints, using the same layout as Figure B-8. $n = 15$ seeds per group. Results are consistent with strict constraints: the optimized prior significantly outperforms all baselines at $N \leq 30$ on CPT MAE and network MAE, the KL advantage over MPL-TC mild is maintained through $N = 100$, and the conflict source comparison remains significant throughout on all three metrics.

B-4 Network-level inference results

B-4-1 Complete Network performance

The main text emphasizes the overall trends in network-level inference performance. The tables in this section provide the complete numerical results underlying those analyses.

Because network queries depend jointly on all CPT rows, these results also serve as a validation that parameter-level improvements propagate to downstream inference tasks.

Table B-5: Network query MAE for mild and strict constraints across all target sizes, averaged over 15 seeds. Baseline columns (Uniform and MPL-TC sources) are identical across constraint modes because those priors are constructed independently of the feasibility constraints. The optimized prior MAE differs between modes, as strict constraints restrict the feasible region further.

Mode	N	Uniform	MPL-TC mild	MPL-TC mod	MPL-TC conflict	Optimized	vs Uni (%)	vs mild (%)	vs mod (%)	vs conflict (%)
Mild	10	0.106	0.090	0.111	0.160	0.036	65.5	59.5	67.3	77.3
	20	0.082	0.081	0.088	0.116	0.033	59.5	59.4	62.6	71.5
	30	0.072	0.059	0.078	0.122	0.032	56.0	46.0	59.4	74.0
	50	0.057	0.040	0.044	0.116	0.033	41.6	17.2	24.9	71.3
	75	0.036	0.025	0.041	0.110	0.025	31.9	1.7	39.8	77.5
	100	0.039	0.031	0.043	0.122	0.022	42.9	28.8	47.5	81.7
	150	0.033	0.020	0.036	0.123	0.020	39.5	0.8	45.3	83.8
	200	0.027	0.021	0.028	0.119	0.018	33.5	13.0	34.5	84.7
Strict	10	0.106	0.090	0.111	0.160	0.040	62.6	56.0	64.5	75.4
	20	0.082	0.081	0.088	0.116	0.037	54.2	54.0	57.7	67.7
	30	0.072	0.059	0.078	0.122	0.032	55.3	45.1	58.7	73.5
	50	0.057	0.040	0.044	0.116	0.034	40.8	16.2	23.9	70.9
	75	0.036	0.025	0.041	0.110	0.027	27.2	−5.1	35.7	75.9
	100	0.039	0.031	0.043	0.122	0.024	38.0	22.7	43.0	80.2
	150	0.033	0.020	0.036	0.123	0.020	37.9	−1.7	43.9	83.3
	200	0.027	0.021	0.028	0.119	0.020	28.6	6.6	29.7	83.6

Table B-6: Network query MAE and mean Bernoulli KL divergence $\mathrm{KL}(p_{\mathsf{true}} \| p_{\mathsf{est}})$ for mild and strict constraints, averaged over 15 seeds and all five queries. Baseline columns are identical across constraint modes.The optimized prior achieves the lowest or near-lowest MAE and KL in most regimes, while consistently outperforming the uniform and conflict baselines; MPL-TC mild yields lower MAE than uniform at small N but higher KL, reflecting that a concentrated but misaligned prior inflates KL more than it inflates MAE.

		MAE					KL divergence				
Mode	N	Uniform	mild	mod	conflict	Opt	Uniform	mild	mod	conflict	Opt
Mild	10	0.106	0.090	0.111	0.160	0.036	0.0331	0.1145	0.1807	0.1298	0.0053
	20	0.082	0.081	0.088	0.116	0.033	0.0239	0.1129	0.0833	0.0457	0.0045
	30	0.072	0.059	0.078	0.122	0.032	0.0198	0.0131	0.0207	0.0519	0.0039
	50	0.057	0.040	0.044	0.116	0.033	0.0130	0.0061	0.0094	0.0442	0.0039
	75	0.036	0.025	0.041	0.110	0.025	0.0045	0.0024	0.0066	0.0358	0.0023
	100	0.039	0.031	0.043	0.122	0.022	0.0056	0.0038	0.0064	0.0425	0.0021
	150	0.033	0.020	0.036	0.123	0.020	0.0040	0.0018	0.0045	0.0428	0.0016
	200	0.027	0.021	0.028	0.119	0.018	0.0034	0.0017	0.0031	0.0385	0.0014
Strict	10	0.106	0.090	0.111	0.160	0.040	0.0331	0.1145	0.1807	0.1298	0.0059
	20	0.082	0.081	0.088	0.116	0.037	0.0239	0.1129	0.0833	0.0457	0.0055
	30	0.072	0.059	0.078	0.122	0.032	0.0198	0.0131	0.0207	0.0519	0.0040
	50	0.057	0.040	0.044	0.116	0.034	0.0130	0.0061	0.0094	0.0442	0.0044
	75	0.036	0.025	0.041	0.110	0.027	0.0045	0.0024	0.0066	0.0358	0.0027
	100	0.039	0.031	0.043	0.122	0.024	0.0056	0.0038	0.0064	0.0425	0.0024
	150	0.033	0.020	0.036	0.123	0.020	0.0040	0.0018	0.0045	0.0428	0.0016
	200	0.027	0.021	0.028	0.119	0.020	0.0034	0.0017	0.0031	0.0385	0.0015

B-4-2 Query-level uncertainty analysis

The main text evaluates network performance using point-estimation metrics such as MAE and KL divergence. To complement these analyses, this section examines the full posterior distributions of the network queries.

The KDE visualizations provide a qualitative view of uncertainty propagation through the Bayesian network. In particular, they show how the optimized priors affect both the location and concentration of the resulting query distributions relative to the uniform and transfer learning baselines.

To verify that the conclusions are not an artifact of KDE-based visualization, query uncertainty was also characterized using the analytical delta-method approximation [63]. The resulting Beta approximations yielded conclusions consistent with the Monte Carlo KDE analysis. The analytical approximation is included as a robustness check and was not used in the main evaluation.

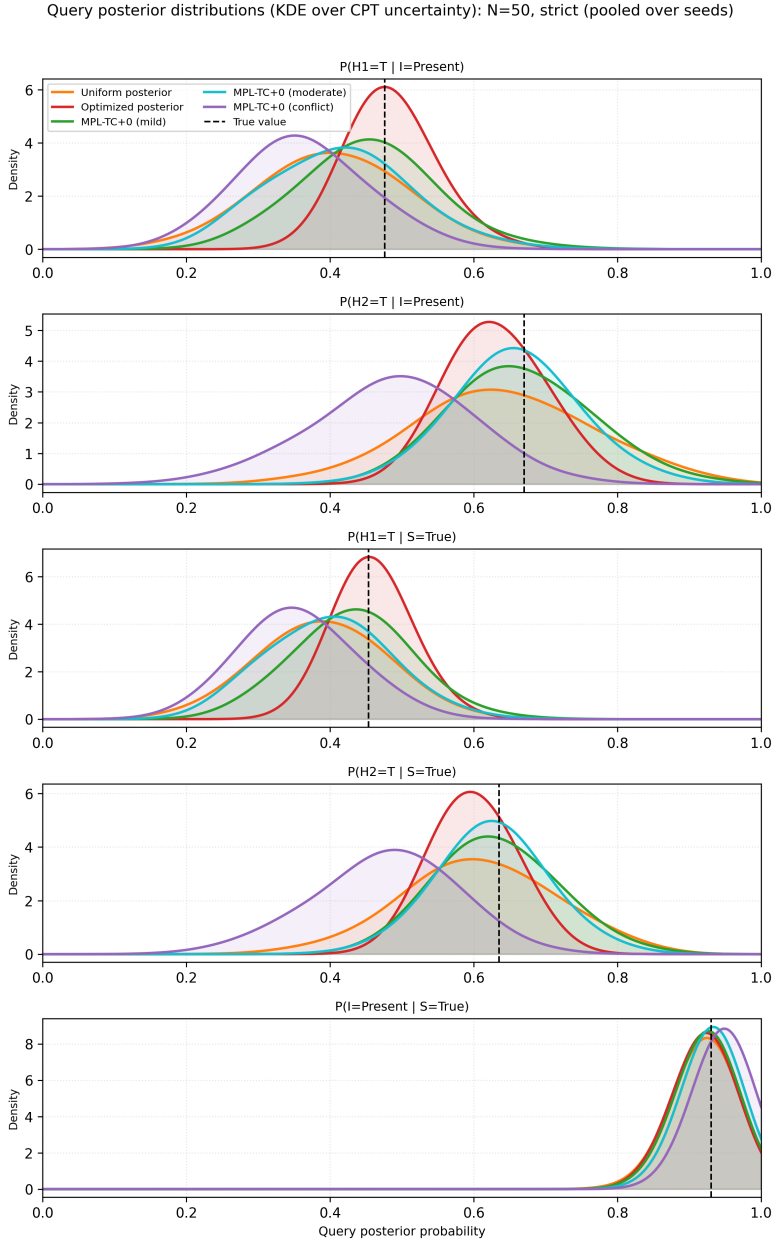


Figure B-10: Full comparision of query posterior distributions represented as KDE curves, pooled over 15 seeds at $N = 50$, strict constraints. Each panel shows one of the five inference queries. The dashed vertical line indicates the true query posterior under the ground-truth BBN. The optimized posterior (red) is more concentrated and its mean is consistently closer to the true value than the uniform posterior (orange) or the MPL-TC source posteriors, reflecting reduced parameter uncertainty and improved distributional placement.

Bibliography

[1] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI Global Scientific Publishing, 2010, pp. 242–264. [Online]. Available: https://doi.org/10.4018/978-1-60566-766-9.ch011

[2] M. Clark, "Bayesian basics: Example," 2025, retrieved October 23, 2025. [Online]. Available: https://m-clark.github.io/bayesian-basics/example.html

[3] Y. Zhou, N. E. Fenton, and M. Neil, "Bayesian network approach to multinomial parameter learning using data and expert judgments," *International Journal of Approximate Reasoning*, vol. 55, no. 5, pp. 1252–1268, 2014. [Online]. Available: https://doi.org/10.1016/j.ijar.2014.02.008

[4] L. Middeldorp, K. Malone, and W. Noordkamp, "Quantifying the robustness of a Bayesian belief network in the context of unmanned aerial system threat prediction," *The Journal of Defense Modeling and Simulation*, p. 15485129231206825, 2023. [Online]. Available: https://dx.doi.org/10.1177/15485129231206825

[5] B. G. Marcot and T. D. Penman, "Advances in bayesian network modelling: Integration of modelling technologies," *Environmental Modelling Software*, vol. 111, pp. 386–393, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1364815218302937

[6] T. Bodewes and M. Scutari, "Learning Bayesian networks from incomplete data with the node-average likelihood," *International Journal of Approximate Reasoning*, vol. 138, pp. 145–160, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0888613X21001328

[7] S. P. van Gosliga and I. Voorde, "Hypothesis management framework: a exible design pattern for belief networks in decision support systems," *CEUR Workshop Proceedings*, vol. 406, 08 2009.

[8] Y. Zhou, N. Fenton, T. Hospedales, and M. Neil, "Probabilistic graphical models parameter learning with transferred prior and constraints," in *UAI*, 01 2015. [Online]. Available: https://doi.org/10.13140/RG.2.1.2714.8645

[9] J. Rohmer, "Uncertainties in conditional probability tables of discrete Bayesian belief networks: A comprehensive review," *Eng. Appl. Artif. Intell.*, vol. 88, 2020. [Online]. Available: https://doi.org/10.1016/j.engappai.2019.103384

[10] L. Mkrtchyan, L. Podofillini, and V. Dang, "Bayesian belief networks for human reliability analysis: A review of applications and gaps," *Reliab. Eng. Syst. Saf.*, vol. 139, pp. 1–16, 2015. [Online]. Available: https://doi.org/10.1016/j.ress.2015.02.006

[11] M. Krieg, "A tutorial on Bayesian belief networks," 2001.

[12] J. Kitson *et al.*, "A survey of Bayesian network structure learning," *arXiv preprint arXiv:2301.09665*, 2023. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s10462-022-10351-w.pdf

[13] M. Druzdzel and H. Simon, "Causality in bayesian belief networks," *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, 03 2013. [Online]. Available: https://doi.org/10.1016/B978-1-4832-1451-1.50005-6

[14] R. Luis, L. E. Sucar, and E. F. Morales, "Inductive transfer for learning Bayesian networks," *Machine learning*, vol. 79, no. 1, pp. 227–255, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:15666728

[15] Y. Yang, X. Gao, Z. Guo, and D. Chen, "Learning Bayesian networks using the constrained maximum a posteriori probability method," *Pattern Recognition*, vol. 91, pp. 123–134, 2019. [Online]. Available: https://doi.org/10.1016/j.patcog.2019.02.006

[16] R. Di, P. Wang, C. He, and Z. Guo, "Constrained adjusted maximum a posteriori estimation of bayesian network parameters," *Entropy*, vol. 23, no. 10, 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/10/1283

[17] R. Luis, L. E. Sucar, and E. F. Morales, "Transfer learning for Bayesian networks," in *Advances in Artificial Intelligence – IBERAMIA 2008*, H. Geffner, R. Prada, I. Machado Alexandre, and N. David, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 93–102. [Online]. Available: https://doi.org/10.1007/978-3-540-88309-8_10

[18] F. G. Cozman, "Credal networks," *Artificial intelligence*, vol. 120, no. 2, pp. 199–233, 2000. [Online]. Available: https://doi.org/10.1016/S0004-3702(00)00029-1

[19] H. J. Suermondt, *Explanation in Bayesian belief networks*. Stanford University, 1992. [Online]. Available: https://api.semanticscholar.org/CorpusID:60120976

[20] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009. [Online]. Available: http://mcb111.org/w06/KollerFriedman.pdf

[21] G. F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine learning*, vol. 9, no. 4, pp. 309–347, 1992. [Online]. Available: https://api.semanticscholar.org/CorpusID:6047868

[22] G. Xiaoguang, Y. Yu, and G. Zhigao, "Learning Bayesian networks by constrained Bayesian estimation," *Journal of Systems Engineering and Electronics*, vol. 30, no. 3, pp. 511–524, 2019. [Online]. Available: https://doi.org/10.21629/JSEE.2019.03.09

[23] J. Pearl, "Bayesian networks: A model of self-activated memory for evidential reasoning," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 7, 1985. [Online]. Available: https://escholarship.org/uc/item/0vr7830n

[24] ——, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988. [Online]. Available: http://books.google.de/books?id=AvNID7LyMusC

[25] S. L. Lauritzen and D. J. Spiegelhalter, "Local computations with probabilities on graphical structures and their application to expert systems," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 50, no. 2, pp. 157–194, 1988. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1988.tb01721.x

[26] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine learning*, vol. 20, no. 3, pp. 197–243, 1995. [Online]. Available: https://doi.org/10.1007/BF00994016

[27] W. L. Buntine, "Operations for learning with graphical models," *Journal of artificial intelligence research*, vol. 2, pp. 159–225, 1994. [Online]. Available: https://doi.org/10.1613/jair.62

[28] Y. You, J. Li, and L. Shen, "An effective Bayesian network parameters learning algorithm for autonomous mission decision-making under scarce data," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 3, pp. 549–561, 2019. [Online]. Available: https://doi.org/10.1007/s13042-017-0737-x

[29] C. P. De Campos and Q. Ji, "Improving Bayesian network parameter learning using constraints," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4. [Online]. Available: https://doi.org/10.1109/ICPR.2008.4761287

[30] M. Ramoni and P. Sebastiani, "Bayesian methods," in *Intelligent Data Analysis*. Springer, 2003. [Online]. Available: https://kmi.open.ac.uk/publications/pdf/kmi-98-8.pdf

[31] A. Karimnezhad, P. J.F. Lucas, and A. Parsian, "Constrained parameter estimation with uncertain priors for Bayesian networks," 2017. [Online]. Available: https://doi.org/10.1214/17-EJS1350

[32] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*, 3rd ed. Chapman & Hall/CRC, 2013. [Online]. Available: https://stat.columbia.edu/~gelman/book/

[33] C. B. Owen, *Parameter estimation for the beta distribution*. Brigham Young University, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:115630033

[34] J. Kruschke *et al.*, *Doing Bayesian data analysis*. Elsevier, 2014, vol. 2. [Online]. Available: https://doi.org/10.5964/ejop.v7i4.163

[35] D. R. S. Saputro, P. Widyaningsih, F. Handayani, and N. A. Kurdhi, "Prior and posterior dirichlet distributions on bayesian networks (BNs)," in *AIP Conference Proceedings*, vol. 1827, no. 1. AIP Publishing LLC, 2017, p. 020036. [Online]. Available: https://doi.org/10.1063/1.4979452

[36] F. Llorente, L. Martino, E. Curbelo, J. López-Santiago, and D. Delgado, "On the safe use of prior densities for Bayesian model selection," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 15, no. 1, p. e1595, 2023. [Online]. Available: https://doi.org/10.1002/wics.1595

[37] X. Gao, Z. Guo, H. Ren, Y. Yang, D. Chen, and C. He, "Learning Bayesian network parameters via minimax algorithm," *International Journal of Approximate Reasoning*, vol. 108, pp. 62–75, 2019. [Online]. Available: https://doi.org/10.1016/j.ijar.2019.03.001

[38] A. Gelman, D. Simpson, and M. Betancourt, "The prior can often only be understood in the context of the likelihood," *Entropy*, vol. 19, no. 10, p. 555, 2017. [Online]. Available: https://www.mdpi.com/1099-4300/19/10/555

[39] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946. [Online]. Available: https://doi.org/10.1098/rspa.1946.0056

[40] B. Wisse, S. P. van Gosliga, N. P. van Elst, and A. I. Barros, "Relieving the elicitation burden of Bayesian belief networks." in *BMA*, 2008. [Online]. Available: https://ceur-ws.org/Vol-406/paper2.pdf

[41] S. Renooij, "Probability elicitation for belief networks: issues to consider," *The Knowledge Engineering Review*, vol. 16, no. 3, pp. 255–269, 2001. [Online]. Available: https://doi.org/10.1017/S0269888901000145

[42] Y. Hou, E. Zheng, W. Guo, Q. Xiao, and Z. Xu, "Learning Bayesian network parameters with small data set: A parameter extension under constraints method," *IEEE Access*, vol. 8, pp. 24 979–24 989, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.2971099

[43] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209–212, 1927. [Online]. Available: https://doi.org/10.1080/01621459.1927.10502953

[44] A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," *The American Statistician*, vol. 52, no. 2, pp. 119–126, 1998. [Online]. Available: https://doi.org/10.1080/00031305.1998.10480550

[45] S. Wallis, "Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods," *Journal of quantitative linguistics*, vol. 20, no. 3, pp. 178–208, 2013. [Online]. Available: https://doi.org/10.1080/09296174.2013.799918

[46] M. J. Druzdel and L. C. Van Der Gaag, *Building probabilistic networks: Where do the numbers come from? - a guide to the literature*, uu-cs ed. Utrecht University: Information and Computing Sciences, 2000, vol. 2000-20. [Online]. Available: http://www.cs.uu.nl/research/techreps/UU-CS-2000-20.html

[47] M. J. Druzdzel, *Probabilistic reasoning in decision support systems: from computation to common sense*. Carnegie Mellon University, 1993. [Online]. Available: https://www.proquest.com/openview/ad4cd1f9a8ba1c05e194eeed6416d087/1?pq-origsite=gscholar&cbl=18750&diss=y

[48] W. Liao and Q. Ji, "Learning Bayesian network parameters under incomplete data with domain knowledge," *Pattern Recognition*, vol. 42, no. 11, pp. 3046–3056, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320309001472

[49] R. S. Niculescu, *Exploiting parameter domain knowledge for learning in bayesian networks*. Carnegie Mellon University, 2005. [Online]. Available: http://reports-archive.adm.cs.cmu.edu/anon/2005/CMU-CS-05-147.pdf

[50] F. Wittig and A. Jameson, "Exploiting qualitative knowledge in the learning of conditional probabilities of Bayesian networks," *arXiv (Cornell University)*, pp. 644–652, jan 2013. [Online]. Available: http://arxiv.org/abs/1301.3903

[51] S. Renooij, L. C. Van der Gaag, and S. Parsons, "Context-specific sign-propagation in qualitative probabilistic networks," *Artificial Intelligence*, vol. 140, no. 1-2, pp. 207–230, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370202002473/pdf?md5=571f2a0d5f9bdf8c212eb373bceda243&pid=1-s2.0-S0004370202002473-main.pdf&_valck=1

[52] Y. Zhou, N. Fenton, and M. Neil, "An extended MPL-C model for Bayesian network parameter learning with exterior constraints," in *European Workshop on Probabilistic Graphical Models*. Springer, 2014, pp. 581–596. [Online]. Available: https://doi.org/10.1007/978-3-319-11433-0_38

[53] R. Chang, M. Stetter, and W. Brauer, "Quantitative inference by qualitative semantic knowledge mining with Bayesian model averaging," *IEEE transactions on knowledge and data engineering*, vol. 20, no. 12, pp. 1587–1600, 2008. [Online]. Available: https://doi.org/10.1109/TKDE.2008.89

[54] R. Chang and W. Wang, "Novel algorithm for Bayesian network parameter learning with informative prior constraints," in *The 2010 international joint conference on neural networks (IJCNN)*. IEEE, 2010, pp. 1–8. [Online]. Available: https://doi.org/10.1109/IJCNN.2010.5596889

[55] R. S. Niculescu, T. M. Mitchell, and R. B. Rao, "A theoretical framework for learning Bayesian networks with parameter inequality constraints." in *IJCAI*, vol. 2, 2007, p. 4. [Online]. Available: https://www.ijcai.org/Proceedings/07/Papers/023.pdf

[56] Y. Zhou, N. Fenton, and C. Zhu, "An empirical study of Bayesian network parameter learning with monotonic influence constraints," *Decision Support Systems*, vol. 87, pp. 69–79, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923616300744

[57] A. Feelders and L. C. Van der Gaag, "Learning Bayesian network parameters under order constraints," *International Journal of Approximate Reasoning*, vol. 42, no. 1-2, pp. 37–53, 2006. [Online]. Available: https://doi.org/10.1016/j.ijar.2005.10.003

[58] Z. Guo, X. Gao, and R. Di, "Learning Bayesian network parameters with domain knowledge and insufficient data," in *Advanced Methodologies for Bayesian Networks*. PMLR, 2017, pp. 93–104. [Online]. Available: https://proceedings.mlr.press/v73/guo17a.html

[59] A. E. Smith, D. W. Coit, T. Baeck, D. Fogel, and Z. Michalewicz, "Penalty functions," *Handbook of evolutionary computation*, vol. 97, no. 1, p. C5, 1997. [Online]. Available: http://140.138.143.31/teachers/Ycliang/Heuristic%20Optimization%20912/penalty%20function.pdf

[60] E. E. Altendorf, "Constraining Bayesian network learning with qualitative models," Master's thesis, Oregon State University, 2005. [Online]. Available: https://ir.library.oregonstate.edu/concern/graduate_thesis_or_dissertations/nc580q460

[61] K. Weichselberger, "The theory of interval-probability as a unifying concept for uncertainty," *International Journal Of Approximate Reasoning*, vol. 24, no. 2–3, pp. 149–170, mei 2000. [Online]. Available: https://doi.org/10.1016/S0888-613X(00)00032-3

[62] Y. Sale, M. Caprio, and E. Höllermeier, "Is the volume of a credal set a good measure for epistemic uncertainty?" in *Uncertainty in Artificial Intelligence*. PMLR, 2023, pp. 1795–1804. [Online]. Available: https://doi.org/10.48550/arXiv.2306.09586

[63] T. Van Allen, A. Singh, R. Greiner, and P. Hooper, "Quantifying the uncertainty of a belief net response: Bayesian error-bars for belief net inference," *Artificial Intelligence*, vol. 172, no. 4–5, pp. 483–513, 2008. [Online]. Available: https://doi.org/10.1016/j.artint.2007.09.004

[64] B. Thijssen and L. F. A. Wessels, "Approximating multivariate posterior distribution functions from monte carlo samples for sequential bayesian inference," *PLOS ONE*, vol. 15, no. 3, p. e0227432, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1712.04200

[65] S. P. Van Gosliga and B. W. Wisse, "Kwantitatieve hypothesevorming: Hypothesis management framework," TNO, The Hague, Report, June 2010.

[66] W. Noordkamp and B. Van Der Vecht, "Hypothesis management framework," TNO, The Hague, Memorandum Report, 2012.

[67] E. Castillo, J. Gutierrez, and A. Hadi, "Sensitivity analysis in discrete Bayesian networks," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 27, no. 4, pp. 412–423, 1997. [Online]. Available: https://doi.org/10.1109/3468.594909

[68] C. P. Robert, N. Chopin, and J. Rousseau, "Harold Jeffreys's Theory of Probability Revisited," *Statistical Science*, vol. 24, no. 2, pp. 141 – 172, 2009. [Online]. Available: https://doi.org/10.1214/09-STS284

[69] B. O'Neill, "Mathematical properties and finite-population correction for the Wilson score interval," *arXiv preprint arXiv:2109.12464*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2109.12464

[70] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004. [Online]. Available: https://doi.org/10.1017/CBO9780511804441

[71] M. Evans and H. Moshonov, "Checking for prior-data conflict," *Bayesian Analysis*, vol. 1, no. 4, pp. 893–914, 2006. [Online]. Available: URL={https://doi.org/10.1214/06-BA129}

[72] A. J. Kleywegt, A. Shapiro, and T. Homem-de Mello, "The sample average approximation method for stochastic discrete optimization," *SIAM Journal on optimization*, vol. 12, no. 2, pp. 479–502, 2002. [Online]. Available: https://optimization-online.org/?p=9206

[73] T. Homem-de Mello and G. Bayraksan, "Monte Carlo sampling-based methods for stochastic optimization," *Surveys in Operations Research and Management Science*, vol. 19, no. 1, pp. 56–85, 2014. [Online]. Available: https://doi.org/10.1016/j.sorms.2014.05.001

[74] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *Journal of applied statistics*, vol. 31, no. 7, pp. 799–815, 2004. [Online]. Available: http://dx.doi.org/10.1080/0266476042000214501

[75] C. Graff, "Expressing relative differences (in percent) by the difference of natural logarithms," *Journal of Mathematical Psychology*, vol. 60, pp. 82–85, 2014.

[76] A. Shapiro, "Monte Carlo sampling approach to stochastic programming," in *ESAIM: proceedings*, vol. 13. EDP Sciences, 2003, pp. 65–73. [Online]. Available: https://doi.org/10.1051/proc:2003003

[77] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., ser. Springer Series in Operations Research and Financial Engineering. New York: Springer, 2006. [Online]. Available: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+502988711&sourceid=fbw_bibsonomy

[78] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965. [Online]. Available: https://doi.org/10.1093/comjnl/7.4.308

[79] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, no. 1, pp. 1–30, 2006. [Online]. Available: http://jmlr.org/papers/v7/demsar06a.html

[80] B. W. Silverman, "Two books on density estimation," *The Annals of Statistics*, vol. 13, no. 4, pp. 1630–1638, 1985. [Online]. Available: http://www.jstor.org/stable/2241378

Glossary

List of Acronyms

BBN	Bayesian Belief Network
CaMAP	Constrained Adjusted Maximum A Posteriori
CBE	Constrained Bayesian Estimation
CI	Credible Interval
CMAP	Constrained Maximum A Posteriori
CPT	Conditional Probability Table
CPTAgg	MatchCPT fitness with Aggregation fusion
DAG	Directed Acyclic Graph
HMF	Hypothesis Management Framework
KDE	Kernel Density Estimation
KL	Kullback Leibler
MAE	Mean Absolute Error
MAP	Maximum A Posteriori
ML	Maximum Likelihood
MPL-C	Multinomial Parameter Learning with Constraints
MPL-TC	Multinomial Parameter Learning with Transferred priors and Constraints
QMAP	Qualitative Maximum A Posteriori
RMSE	Root Mean Square Error