Some New Empirical Analysis of Evaluating Iterative Join-Graph Propagation

Emma Rollon and Rina Dechter
Department of Information and Computer Science
University of California, Irvine
{erollon, dechter}@ics.uci.edu

1 Introduction

In [1], the authors show that IBP [3] (or equivalently, the more general class of algorithm called IJGP [2]) is sound with respect to the inference of zero beliefs. In this report, we empirically investigate the behaviour of IBP/IJGP for near zero inferred beliefs. Specifically, we explore the hypothesis that if IBP infers that the belief of a variable is close to zero (i.e., $\epsilon \to 0$) then this inference is relatively accurate. The study includes some empirical results from [2] and significant new analysis of empirical evaluation carried on in UAI 2006 and UAI 2008 benchmarks[1]. We will see that while our empirical results support the hypothesis on benchmarks having no determinism, the results are quite mixed for networks with determinism.

2 Methodology

We test the accuracy of IBP/IJGP's prediction across the range of belief values from 0 to 1. For IJGP, we also test its performance for increasing values of the control parameter i-bound. We report the results by means of *absolute errors* and *distances* plots:

1. **Absolute Errors**. Using names inspired by the well known measures in information retrieval, we report *Recall Absolute Error* and *Precision Absolute Error* over small intervals spanning $[0, 1]$. *Recall* is the absolute error averaged over all the exact beliefs that fall into the interval, and can therefore be viewed as capturing the level of completeness. For *precision*, the average is taken over all the belief values computed by IBP/IJGP that fall into the interval, and can be viewed as capturing soundness. The left Y axis corresponds to the histograms (the bars), the right Y axis corresponds to the absolute error (the lines).

[1] http://graphmod.ics.uci.edu/uai08/Evaluation/Report

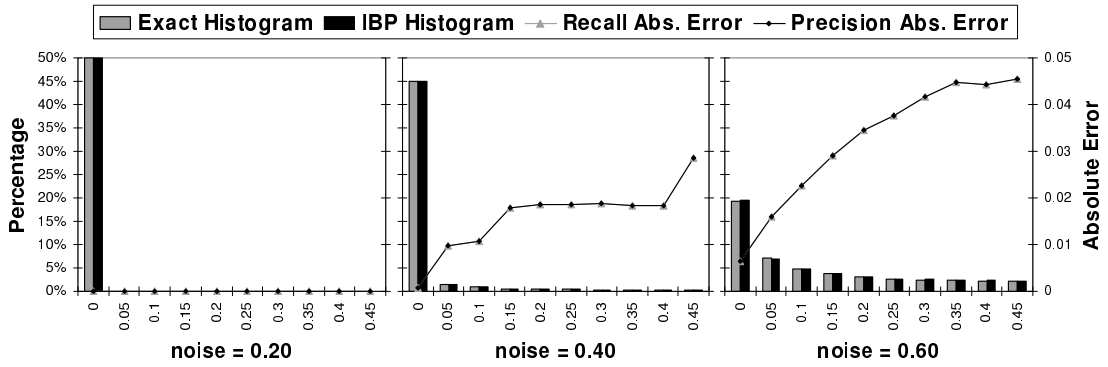

Figure 1: Coding, N=200, evidence=100, w*=15, 1000 instances.

2. **Distances**. For each interval, we distinguish the number of inferred marginals such that (i) their correspoding exact marginals also fall in that interval (i.e., they are '*correct*'); (ii) their correspoding exact marginals fall in a higher interval (i.e., they are '*smaller*'); and (iii) their correspoding exact marginals fall in a smaller interval (i.e., they are '*higher*'). For each set of misplaced inferred marginals (i.e., smaller and higher), we indicate the mean distance to the interval where they belong (i.e., '*dist. up*' and '*dist. down*', respectively). The left Y axis corresponds to the histograms (the bars), the right Y axis corresponds to the distance (the lines).

We discretize the x-axis in intervals of size ϵ. The X coordinate in Figure 1 and Figure 8 denotes the interval $[X, X + \epsilon)$. For the rest of figures, the X coordinate denotes the interval $(X - \epsilon, X]$, where the 0 interval is $[0, 0]$. For problems with binary variables, we only show the interval [0, 0.5] because the graphs are symmetric around 0.5. The number of variables, number of evidence variables and induced width w* are reported in each graph.

We test the performance of IJGP on five different benchmarks: *coding*, *linkage analysis*, *grids*, *two-layer noisy-or*, and *CPCS* networks. Since the behavior within each benchmark is similar, in the following section we report a subset of the results, all of them setting ϵ to 0.05. Complete results (i.e., absolute error and distance analysis setting $\epsilon = 0.05$ for the remaining instances, and absolute and distance analysis setting $\epsilon = 0.005$ for *linkage* and *grids* networks) can be found in Appendix A.

3 Results

Coding networks are the famous case where IBP has impressive performance. The instances are from the class of linear block codes, with 50 nodes per layer and 3 parent nodes for each variable. We experiment with instances having three different values of channel noise: 0.2, 0.4 and 0.6. For each channel value, we generate 1000 samples.

Figure 1 shows the results. When the noise level is 0.2, all the beliefs computed by IBP are extreme. The Recall and Precision are very small, of the order of 10^{-11}. So, in this case, all the beliefs are very small (i.e., ϵ small) and IBP is able to infer them correctly, resulting in almost perfect accuracy (IBP is indeed perfect in this case for the bit error rate). As noise increases, the Recall and Precision get closer to a bell shape, indicating higher error for values close to 0.5 and smaller error for extreme values. The histograms show that fewer belief values are extreme as noise increases.

Linkage Analysis networks. Genetic linkage analysis is a statistical method for mapping genes onto a chromosome. The problem can be modeled as a Bayesian network. We experimented with four *pedigree* instances from the UAI 2008 competition. The domain size ranges between 1 to 4. For these instances exact results are available.

Figure 2 shows the *absolute error* results. We observe that the number of exact 0 beliefs is small and IJGP correctly infers all of them. The behavior of IJGP for ϵ small beliefs varies accross instances. For *pedigree1*, the Exact and IJGP histograms are about the same (for all intervals). Moreover, Recall and Precision errors are relatively small. For the rest of instances, the accuracy of IJGP for extreme inferred marginals decreases. Notice that IJGP infers more ϵ small beliefs than the number of exact extremes in the corresponding intervals, leading to relatively high Precision error while small Recall error. The behaviour for beliefs in the 0.5 interval is reversed, leading to high Recall error while small Precision error. As expected, the accuracy of IJGP improves as the value of the control parameter i-bound increases.

Let us now consider the *distance* results of Figure 3. IJGP misplaces marginals in all intervals different from $[0, 0]$. The number of erroneous marginals is relatively high in all instances, with the exception of *pedigree1*. For i-bound equal to 3 and ϵ small beliefs (i.e., beliefs in interval $(0, 0.05]$), IJGP correctly infers almost all true beliefs that fall in that interval. Note that only a very small percentage of true beliefs that fall in that interval is erroneously inferred in interval $(0.05, 0.1]$. As a consequence, the Recall error is small, as observed in the absolute error results. However, IJGP is very inaccurate on the misplaced marginals. The distance to their correct interval is from 5 up to 8 intervals, depending on the instance. As a consequence, the Precision error is relatively high, as observed in the absolute error results. As the i-bound increases, the number of misplaced marginals, as well as the distance to the correct interval, decreases which leads to a better accuracy.

Grid networks. Grid networks are characterized by two parameters (N, D), where $N \times N$ is the size of the network and D is the percentage of determinism (i.e., the percentage of values in all CPTs assigned to either 0 or 1). We experiment with *grids2* instances from the UAI08 competition. They are characterized by parameters $(\{16, \ldots, 42\}, \{50, 75, 90\})$. For each parameter configuration, there are samples of size 10 generated by randomly assigning value 1 to one leaf node.

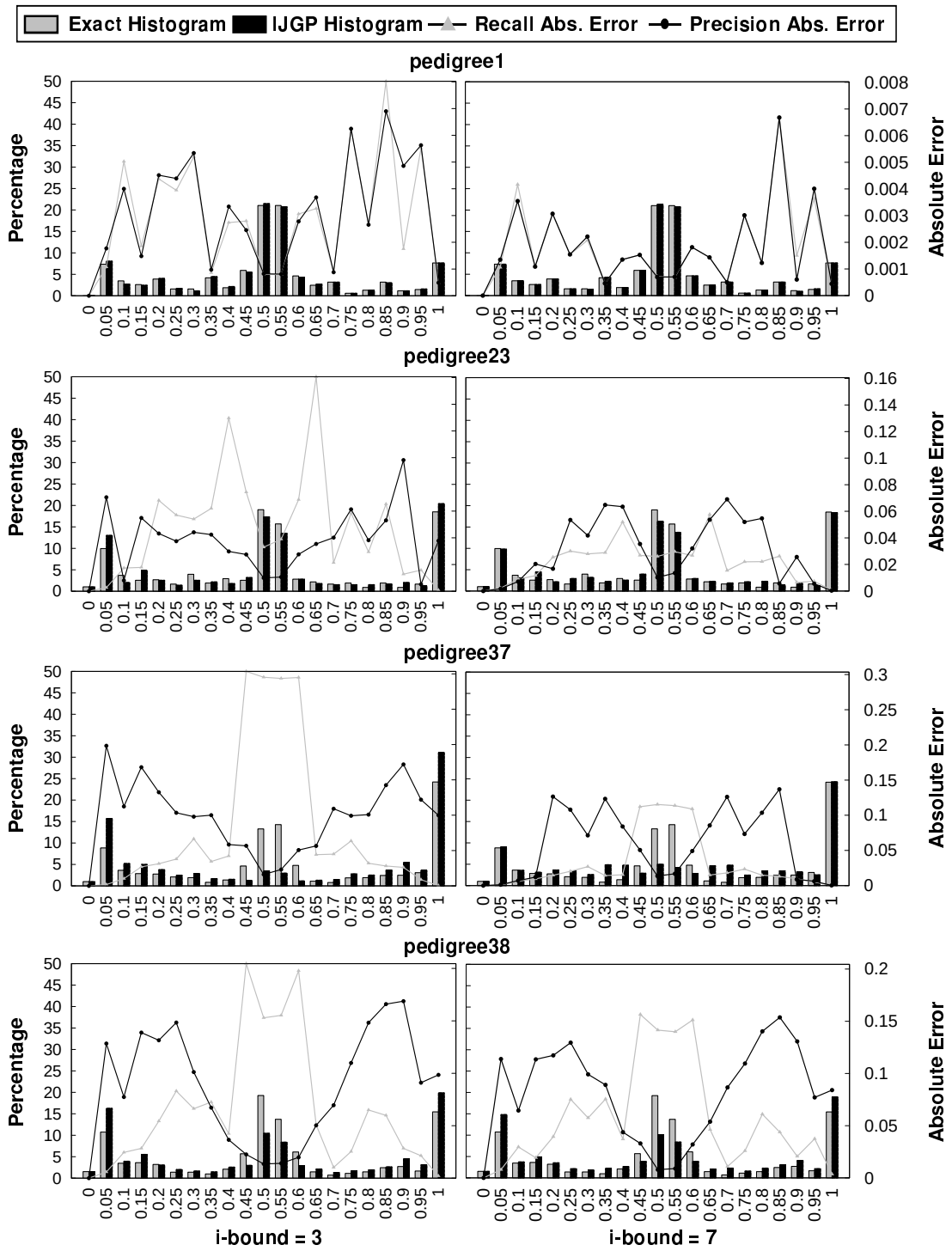

Figure 2: *Absolute error* results on pedigree instances. Each row is the result for one instance. Each column is the result of running IJGP with i-bound equal to 3 and 7, respectively. The number of variables N, number of evidence variables NE, and induced width w* of each instance is as follows. Pedigree1: $N = 334$, $NE = 36$ and w*=21; pedigree23: $N = 402$, $NE = 93$ and w*=30; pedigree37: $N = 1032$, $NE = 306$ and w*=30; pedigree38: $N = 724$, $NE = 143$ and w*=18.

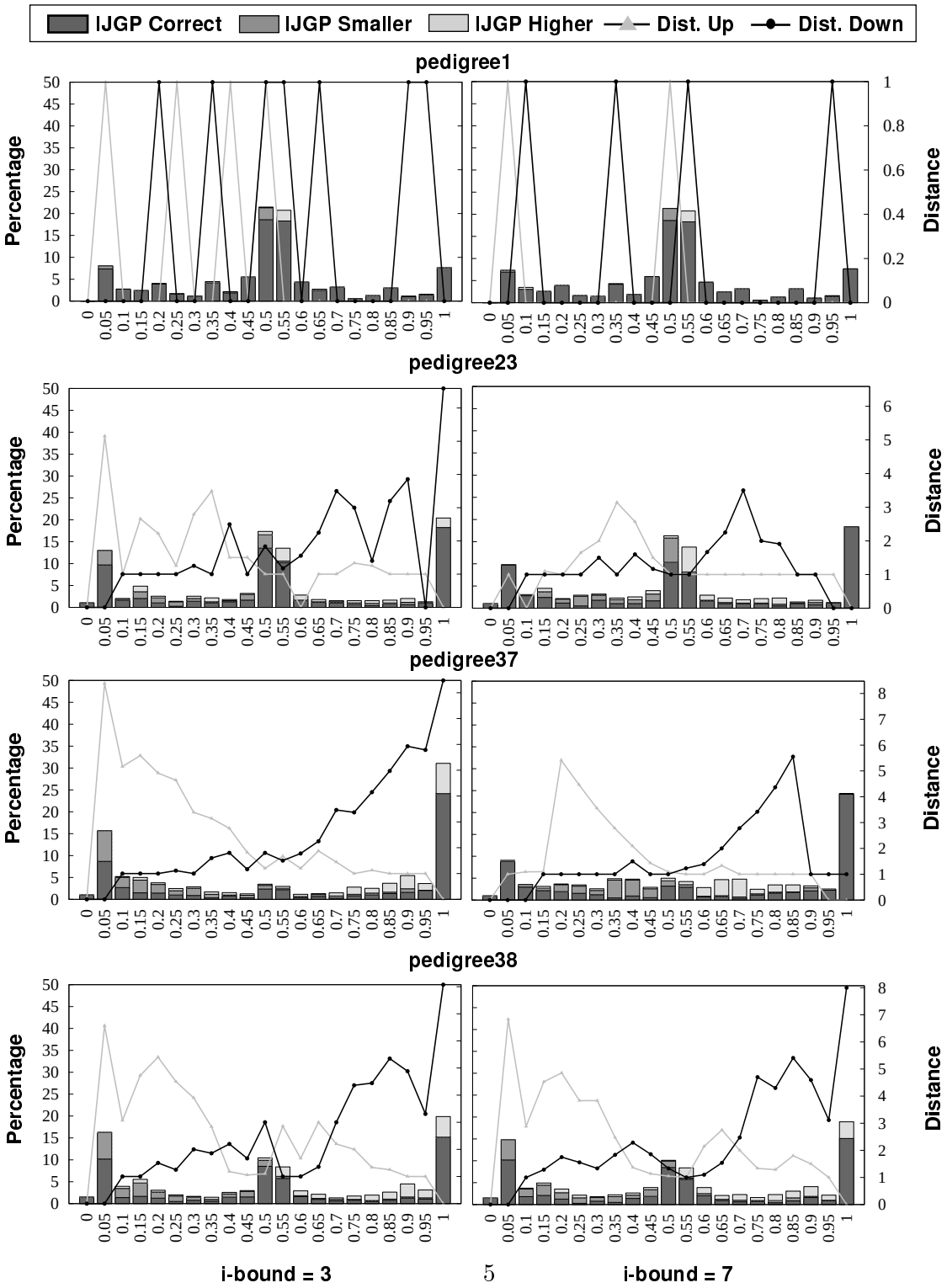

Figure 3: *Distance* results on pedigree instances. Each row is the result for one instance. Each column is the result of running IJGP with i-bound equal to 3 and 7, respectively.

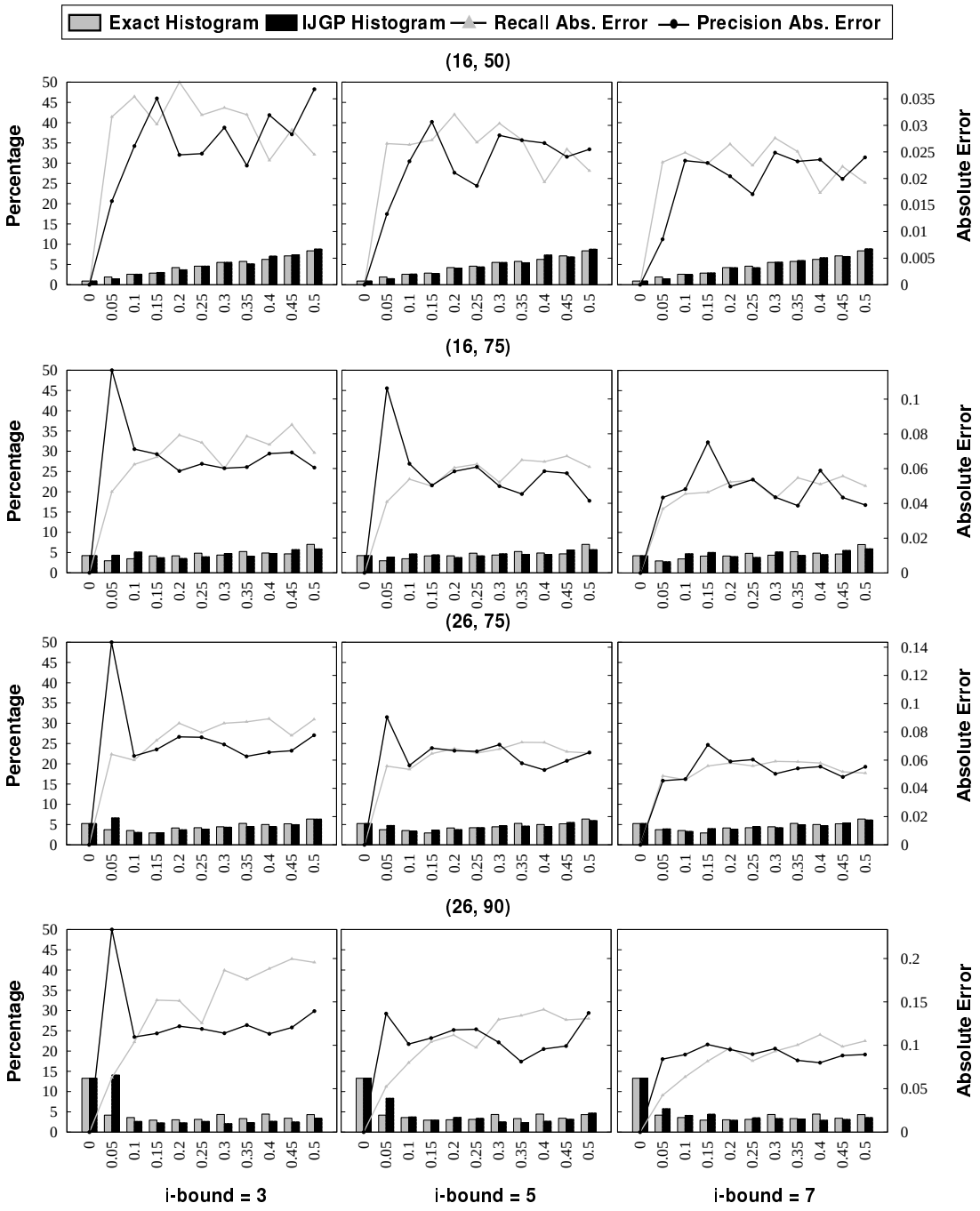

Figure 4: *Absolute error* results on grids2 instances. First and second rows show the results for parameter configuration $(16, 50)$ and $(16, 75)$, respectively. Third and fourth two rows show the results for $(26, 75)$ and $(26, 90)$, respectively. Each column is the result of running IJGP with i-bound equal to 3, 5, and 7, respectively. Each plot indicates the mean value for up to 10 instances. All parameter configurations have $N \times N$ variables and one evidence variable. Configuration $(16, *)$ has induced width w*=22, while $(26, *)$ has w*=40.

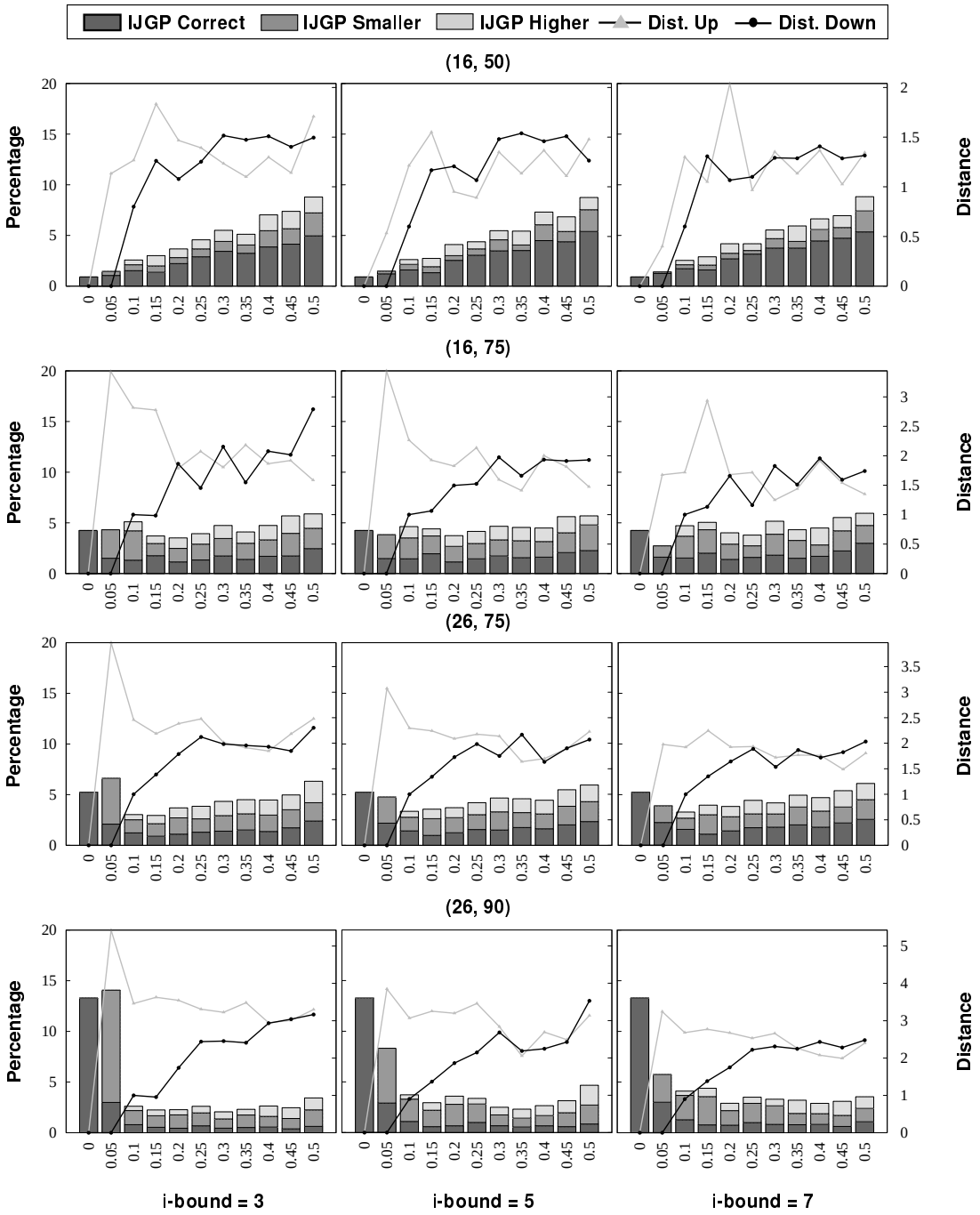

Figure 5: *Distance* results on grids2 instances. First and second rows show the results for parameter configuration $(16, 50)$ and $(16, 75)$, respectively. Third and fourth two rows show the results for $(26, 75)$ and $(26, 90)$, respectively. Each column is the result of running IJGP with i-bound equal to 3, 5, and 7, respectively. Each plot indicates the mean value for up to 10 instances.

Figure 4 reports the *absolute error* results. IJGP correctly infers all 0 beliefs. However, its performance for ϵ small beliefs is quite poor. Only for networks with parameters $(16, 50)$ the Precision error is relatively small (less than 0.05). If we fix the size of the network and the i-bound, both Precision and Recall errors increase as the determinism level D increases. The histograms clearly show the gap between the number of true ϵ small beliefs and the ones inferred by IJGP. As before, the accuracy of IJGP improves as the value of the control parameter i-bound increases. However, note that the improvement is quite significant for interval $(0, 0.05]$ while quite small for the other intervals.

Figure 5 reports the *distance* results. In general, the number of misplaced marginals in all intervals different from 0 is very high. For i-bound equal to 3, the percentage of correctly inferred true ϵ small beliefs is very small with respect to the number of misplaced beliefs. Note that the increase in the value of the i-bound clearly reduces the number of misplaced ϵ small marginals. However, as observed in the previous figures, its effect is less important in the other intervals.

Two-layer noisy-OR networks. Variables are organized in two layers where the ones in the second layer have 10 parents. Each probability table represents a noisy OR-function. Each parent variable y_j has a value $P_j \in [0..P_{noise}]$. The CPT for each variable in the second layer is then defined as, $P(x = 0|y_1, \ldots, y_P) = \prod_{y_j=1} P_j$ and $P(x = 1|y_1, \ldots, y_P) = 1 - P(x = 0|y_1, \ldots, y_P)$. We experiment on *bn2o* instances from the UAI08 competition.

Figure 6 reports the *absolute error* results for 3 instances. In this case, IJGP is very accurate for all instances. In particular, the accuracy in ϵ small beliefs is very high. The *distance* results in Figure 7 show that all inferred marginals are placed in the correct interval. The only exception is for instances *bn2o-30-20-200-1a* and *bn2o-30-25-250-1a*, for which a very small percentage of inferred marginals is misplaced. However, note that they belong to the neighbour interval (i.e., their distance to the correct interval is 1).

CPCS networks. These are medical diagnosis networks derived from the Computer-Based Patient Care Simulation system (CPCS) expert system. We tested on two networks, *cpcs54* and *cpcs360*, with 54 and 360 variables, respectively. For the first network, we generate samples of size 100 by randomly assigning 10 variables as evidence. For the second network, we also generate samples of the same size by randomly assigning 20 and 30 variables as evidence.

Figure 8 shows the *absolute error* results. The histograms show opposing trends in the distribution of beliefs. Although irregular, the absolute error tends to increase towards 0.5 for cpcs54. In general, the error is quite small throughout all intervals and, in particular, for inferred extreme marginals.

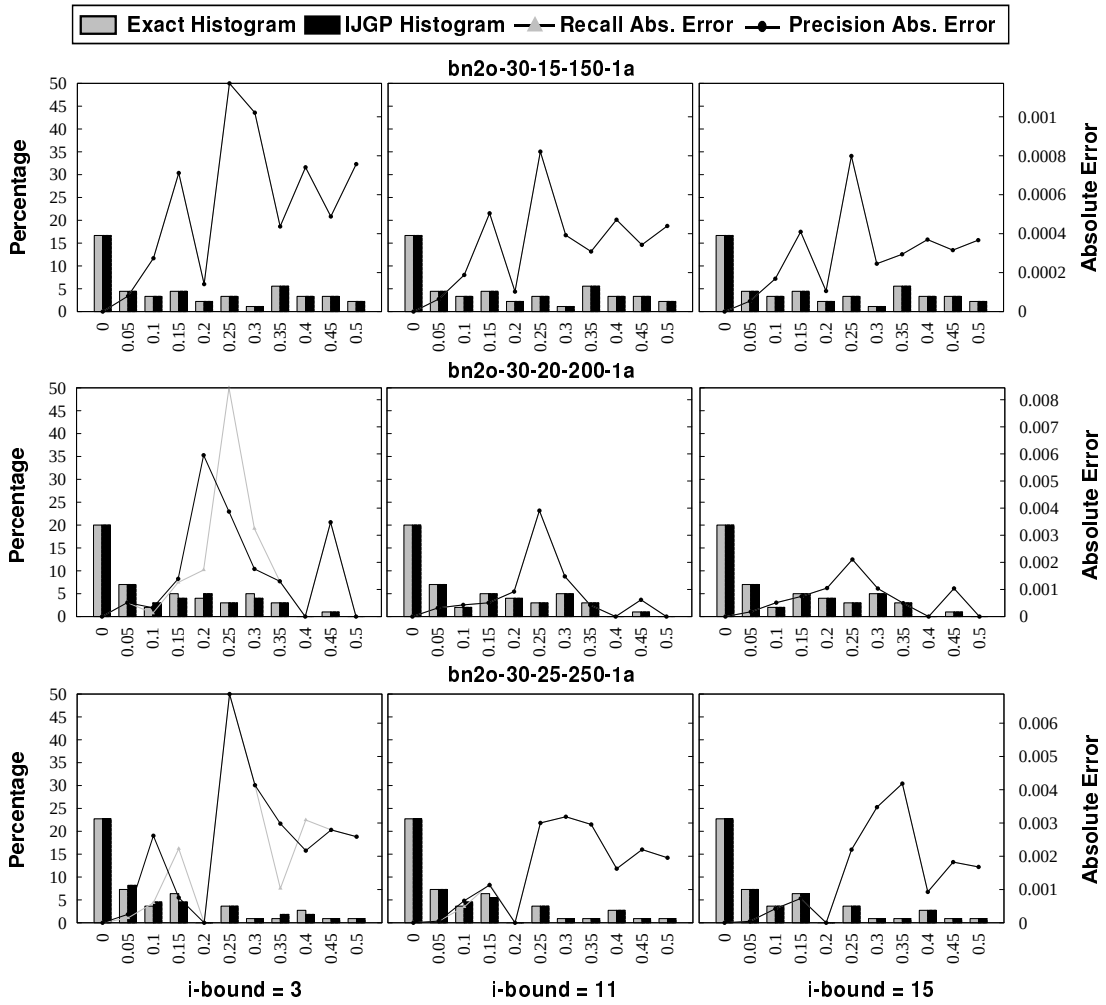

Figure 6: *Absolute error* results on bn2o instances. Each row is the result for one instance. Each column in each row is the result of running IJGP with i-bound equal to 3, 5 and 7, respectively. The number of variables N, number of evidence variables NE, and induced width w* of each instance is as follows. bn2o-30-15-150-1a: $N = 45$, $NE = 15$, and w*=24; bn2o-30-20-200-1a: $N = 50$, $NE = 20$, and w*=27; bn2o-30-25-250-1a: $N = 55$, $NE = 25$, and w*=26.

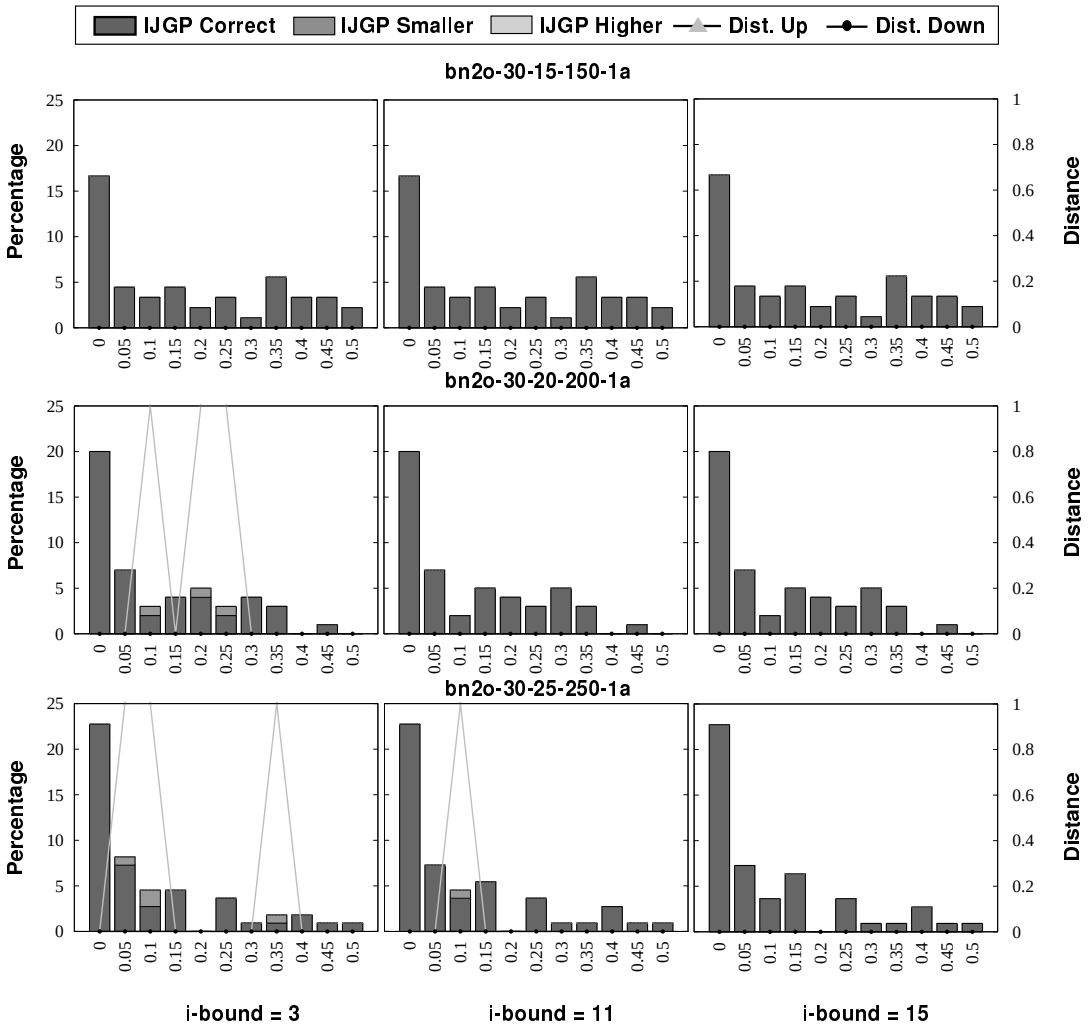

Figure 7: *Distance* results on bn2o instances. Each row is the result for one instance. Each column in each row is the result of running IJGP with i-bound equal to 3, 5 and 7, respectively.

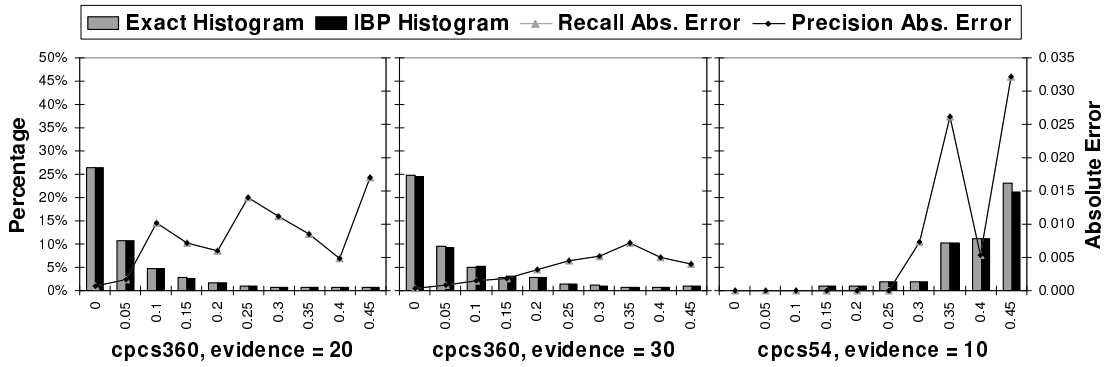

Figure 8: CPCS54, 100 instances, w*=15; CPCS360, 5 instances, w*=20

4 Discussion

The results presented in this report validate that (i) IBP/IJGP inferred zeros are sound; and, (ii) IBP/IJGP is not always sound when inferring ϵ small beliefs. Based on the empirical work, the accuracy of belief propagation in its varied strength levels inferring extreme beliefs seems to depend, in part, on the level of determinism. For instances without determinism, BP's inference near zero was sound in the sense that the average absolute error as expressed by precision was contained within the length of the 0.05 interval (see *two layer noisy-OR* and *CPCS* benchmarks). However, IBP exhibits mixed behavior near zero in the presence of determinism. The experiments on *coding* networks show that IBP is almost perfect. However, for *pedigree* and *grid* networks the results are inaccurate near zeros.

References

[1] R. Dechter and R. Mateescu. A simple insight into iterative belief propagation's success. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI'03)*, pages 175–183, 2003.

[2] R. Dechter, R. Mateescu, and K. Kask. Iterative join-graph propagation. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI'02)*, pages 128–136, 2002.

[3] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann Publishers, 1988.

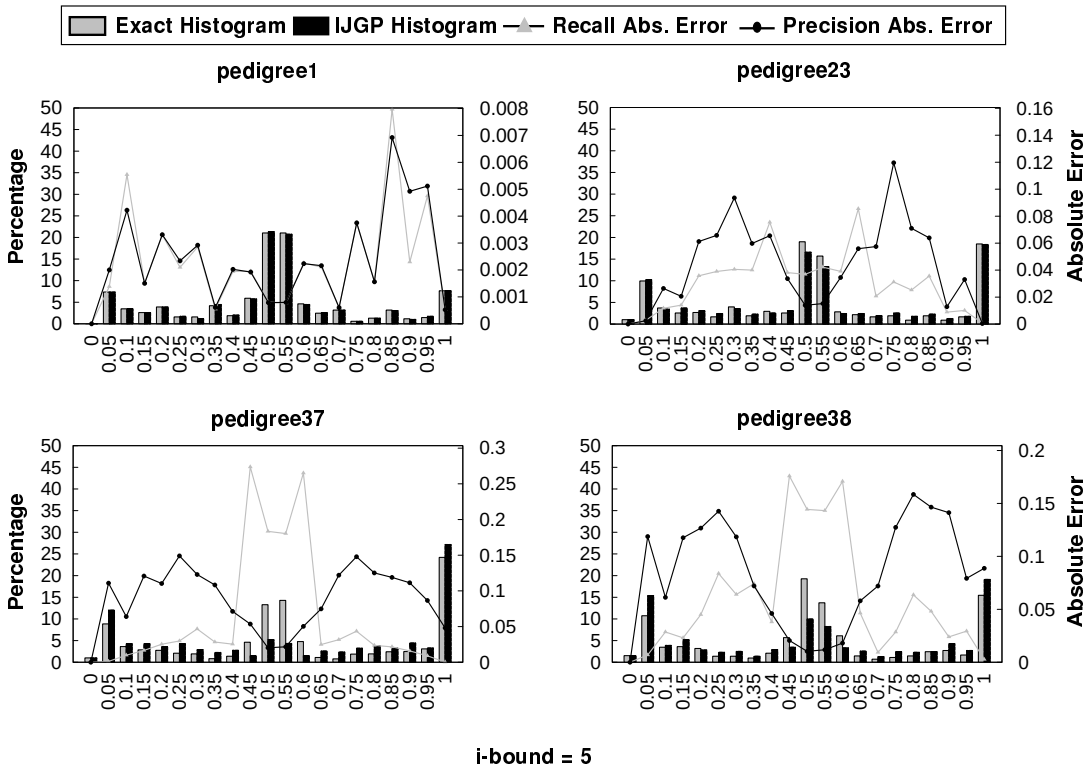

Figure 9: *Absolute error* results on pedigree instances. $\epsilon = 0.05$.

A Complete Results

The next figures (i.e., Figure 9 to Figure 35) show the *absolute error* and *distance* results for the subset of instances not reported in Section 3. We only report *distance* results for instances having some misplaced marginals.

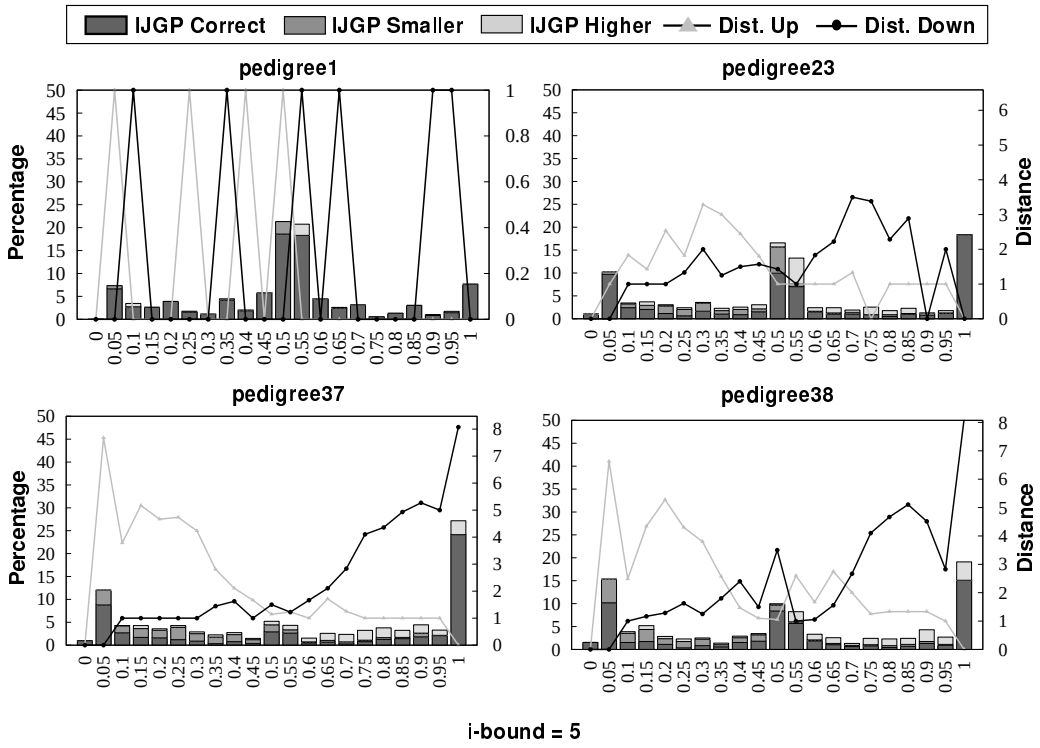

Figure 10: *Distance* results on pedigree instances. $\epsilon = 0.05$.

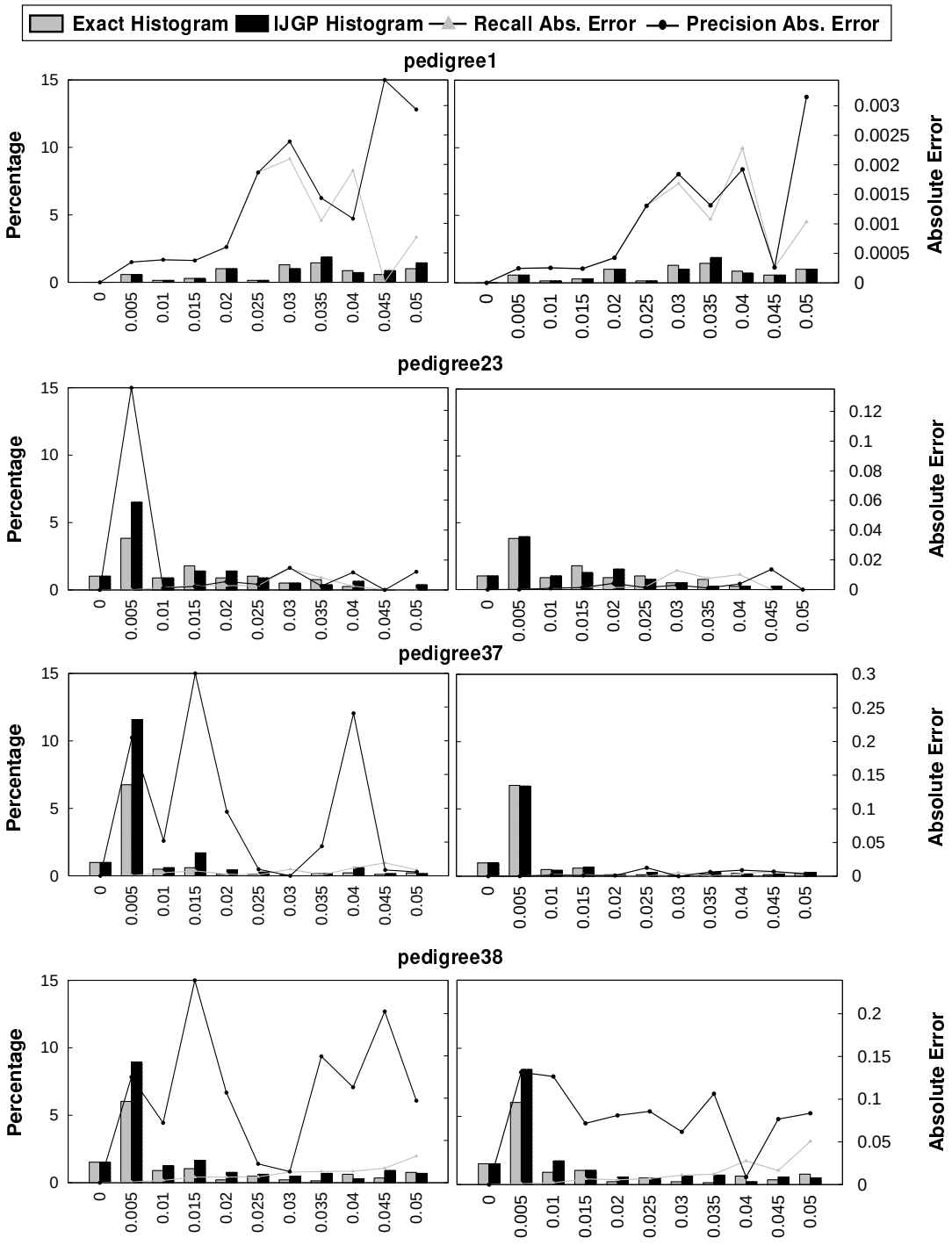

Figure 11: *Absolute error* results on pedigree instances. $\epsilon = 0.005$.

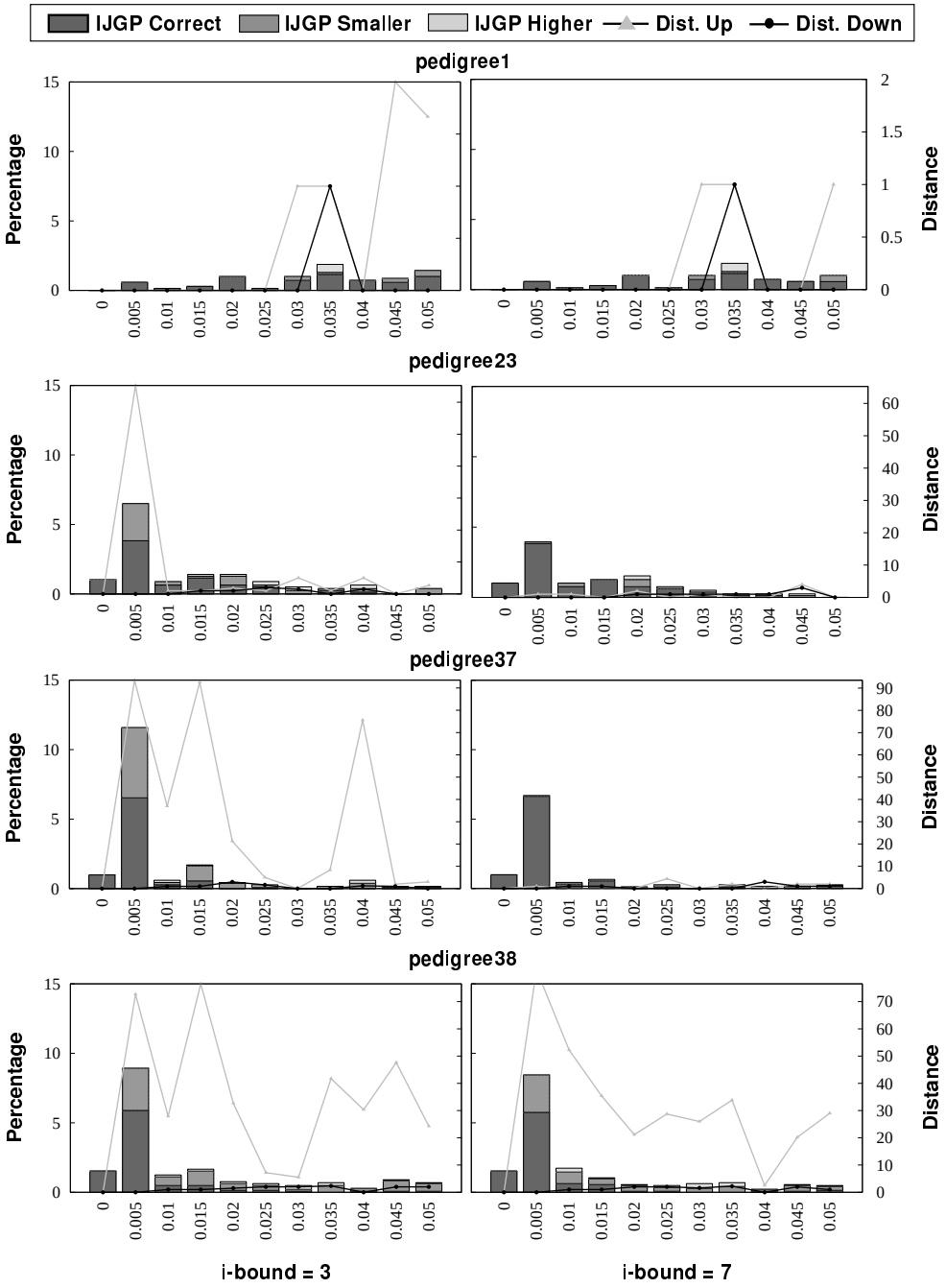

Figure 12: *Distance* results on pedigree instances. $\epsilon = 0.005$.

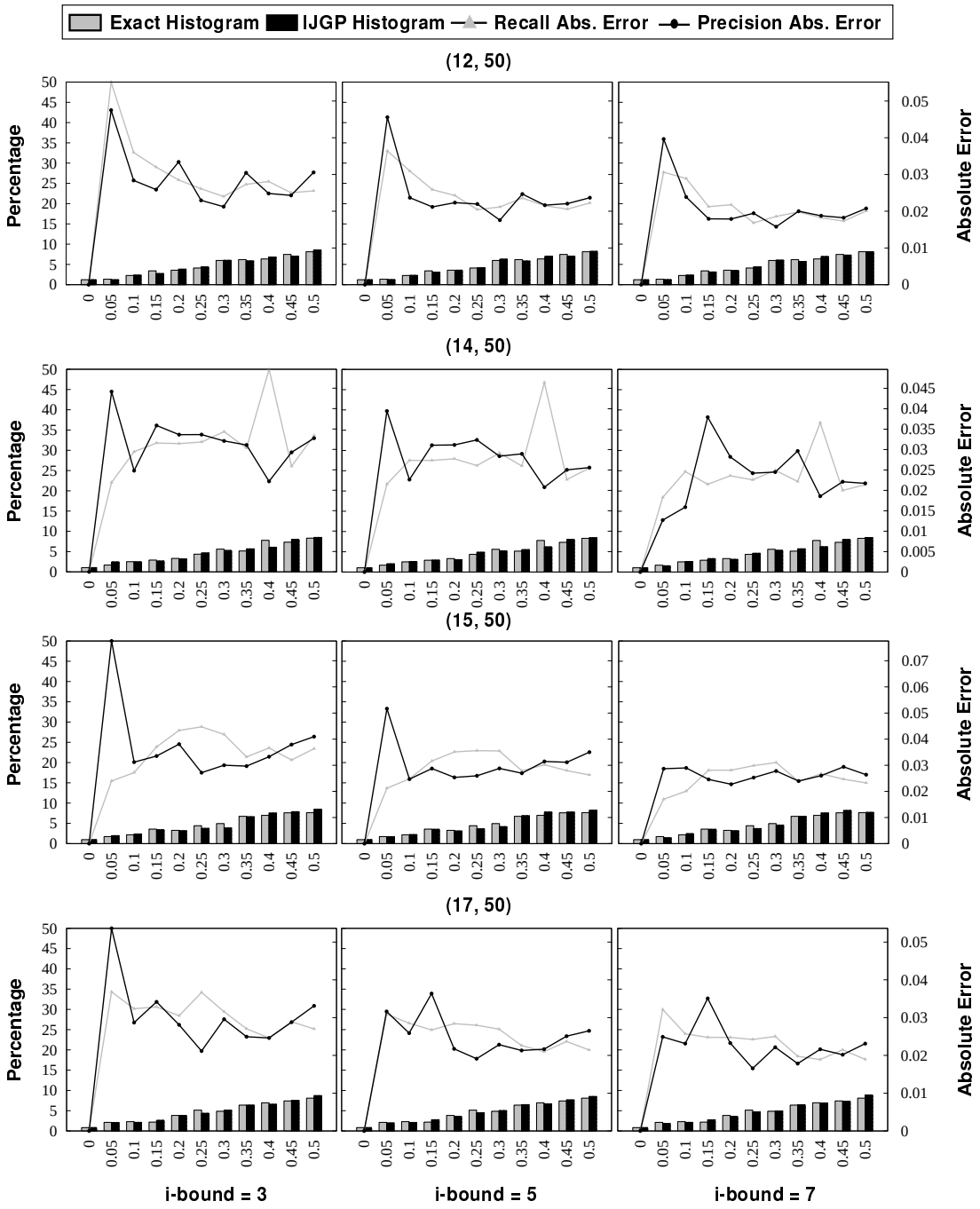

Figure 13: *Absolute error* results on grids instances. Each row shows the results for one parameter configuration (N, D). Each parameter configuration has $N \times N$ variables and one evidence node. The induced width w* is 16, 20, 21 and 24, respectively. $\epsilon = 0.05$

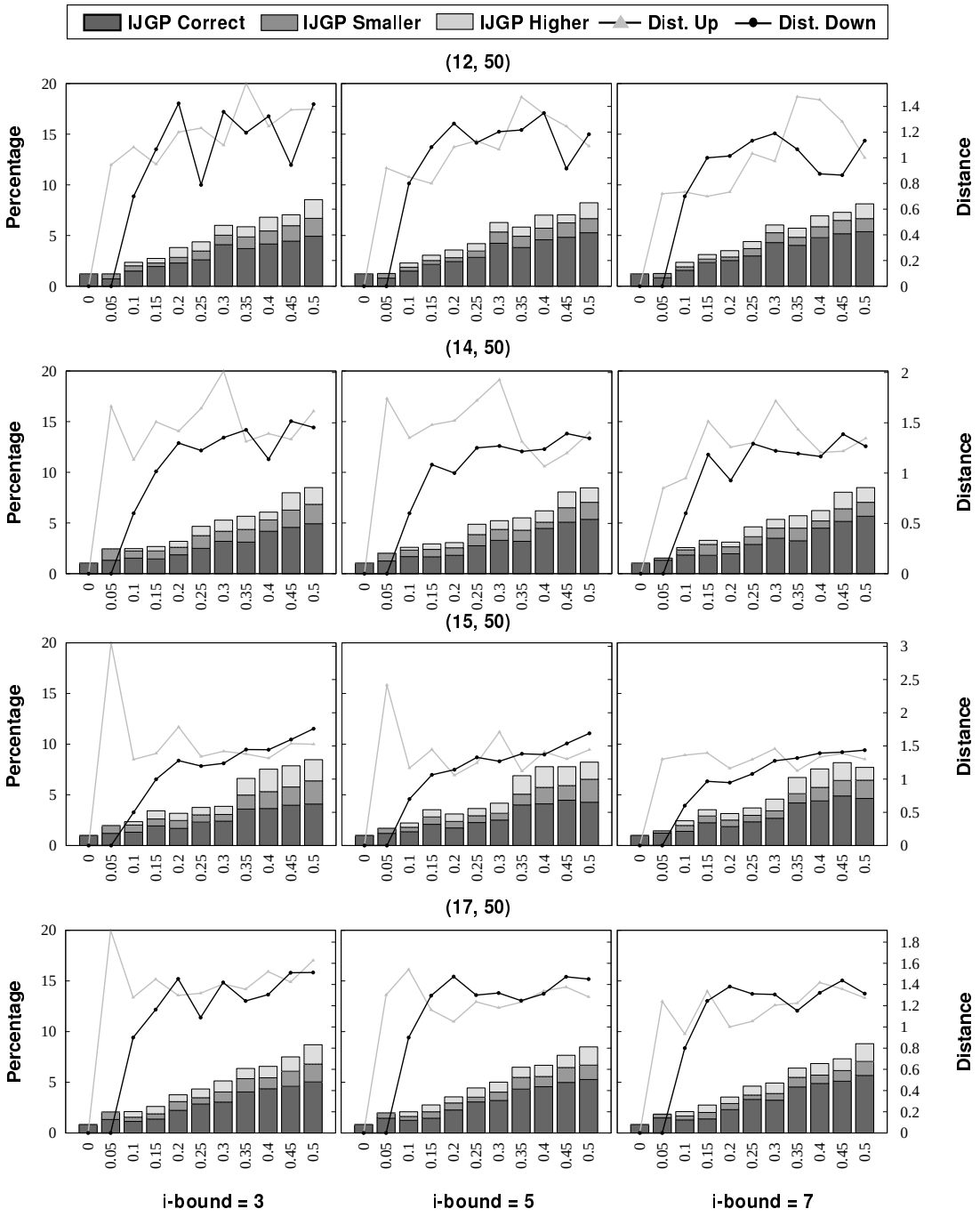

Figure 14: *Distance* results on grids instances. $\epsilon = 0.05$

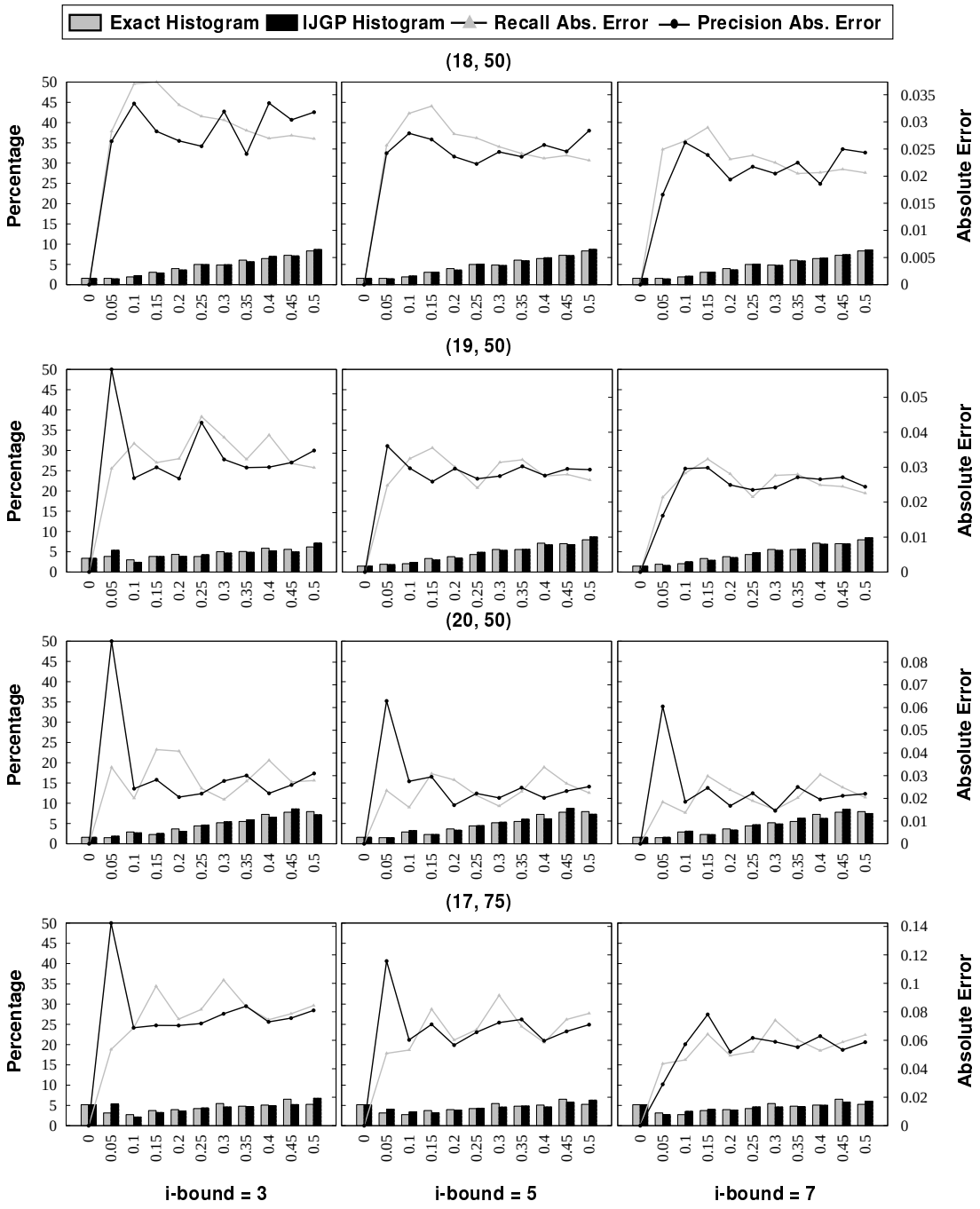

Figure 15: *Absolute error* results on grids instances. Each row shows the results for one parameter configuration (N, D). Each parameter configuration has $N \times N$ variables and one evidence node. The induced width w* is 26, 28, 29 and 24, respectively. $\epsilon = 0.05$

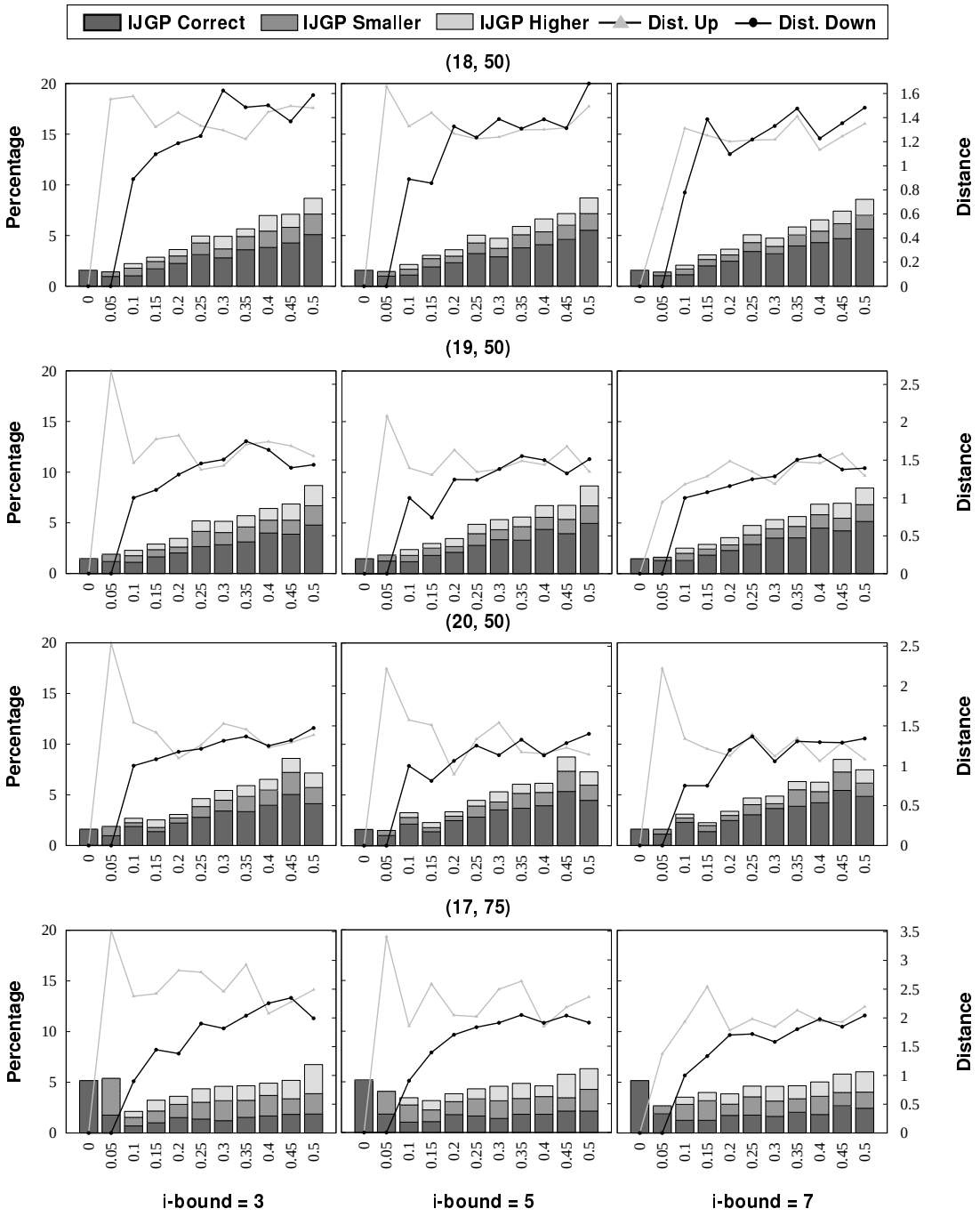

Figure 16: *Distance* results on grids instances. $\epsilon = 0.05$

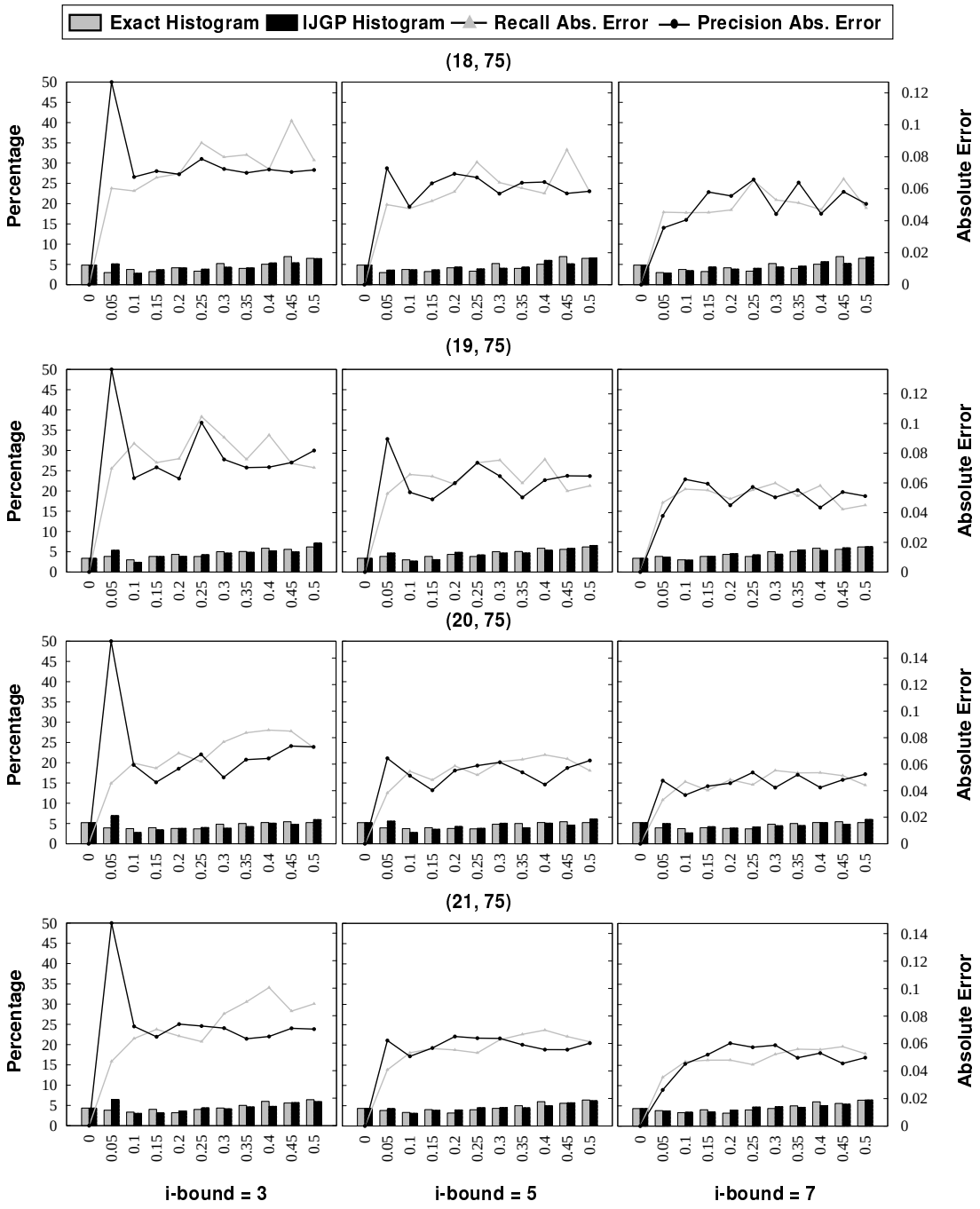

Figure 17: *Absolute error* results on grids instances. Each row shows the results for one parameter configuration (N, D). Each parameter configuration has $N \times N$ variables and one evidence node. The induced width w* is 26, 28, 29 and 31, respectively. $\epsilon = 0.05$

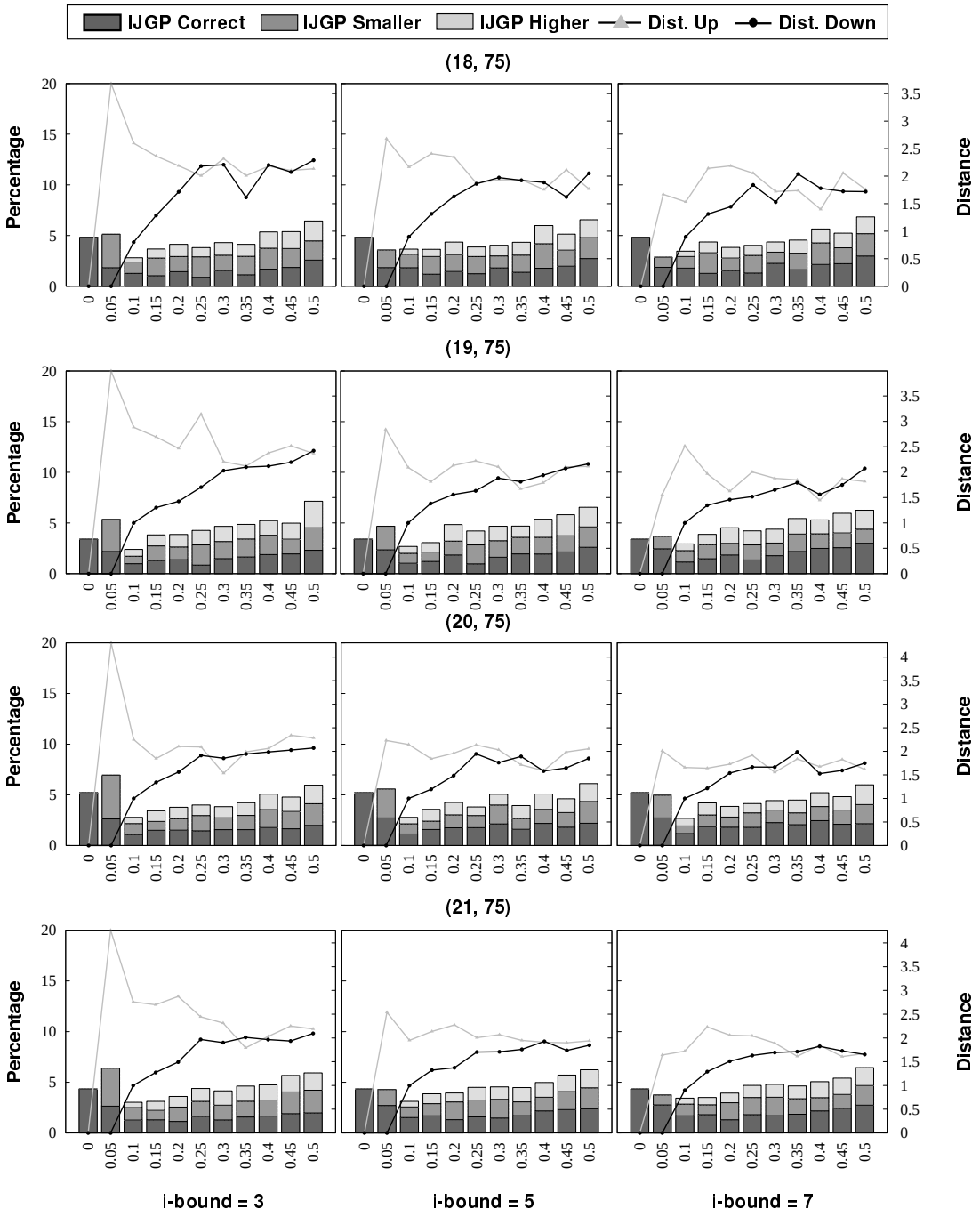

Figure 18: *Distance* results on grids instances. $\epsilon = 0.05$

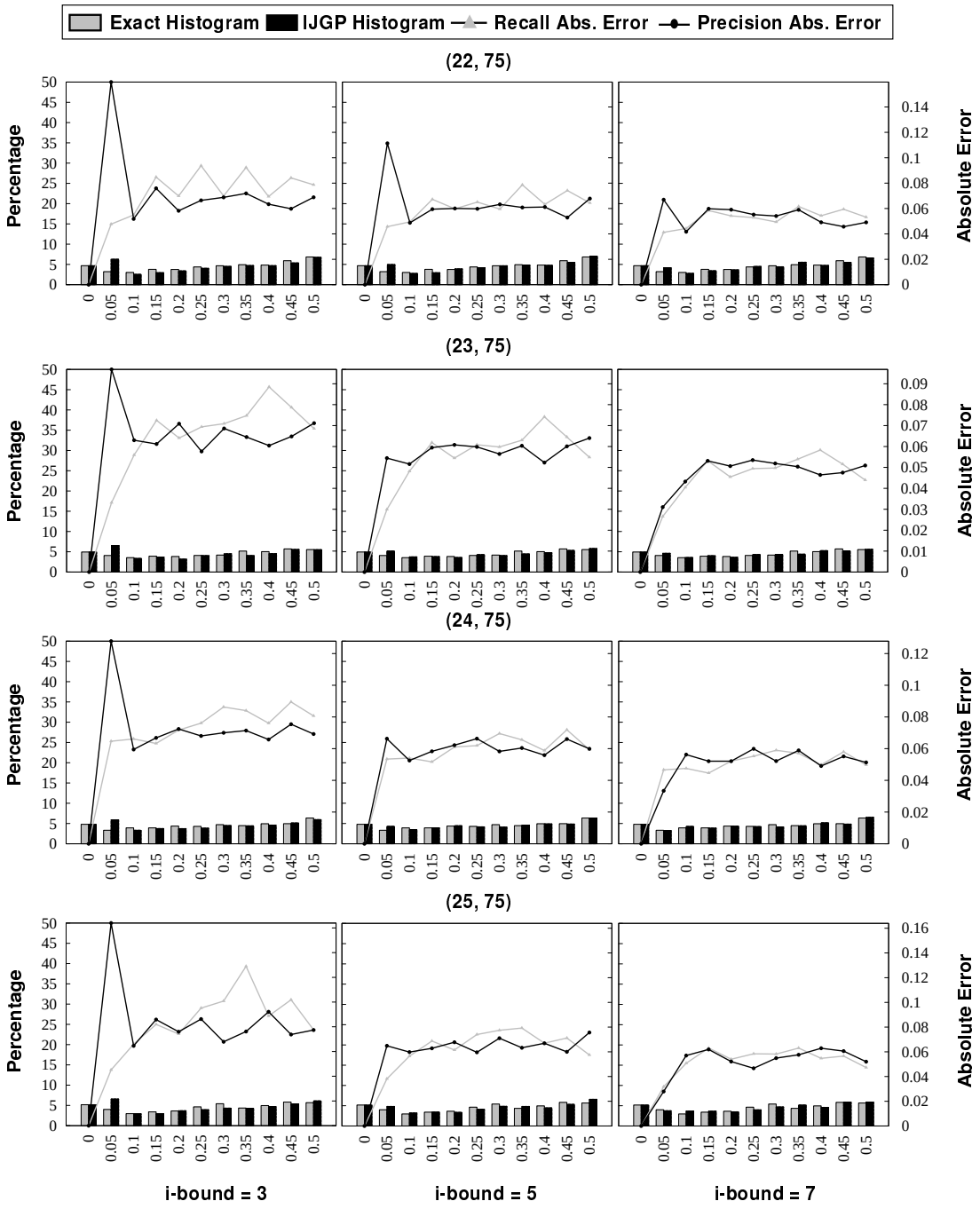

Figure 19: *Absolute error* results on grids instances. Each row shows the results for one parameter configuration (N, D). Each parameter configuration has $N \times N$ variables and one evidence node. The induced width w* is 32, 33, 36 and 38, respectively. $\epsilon = 0.05$

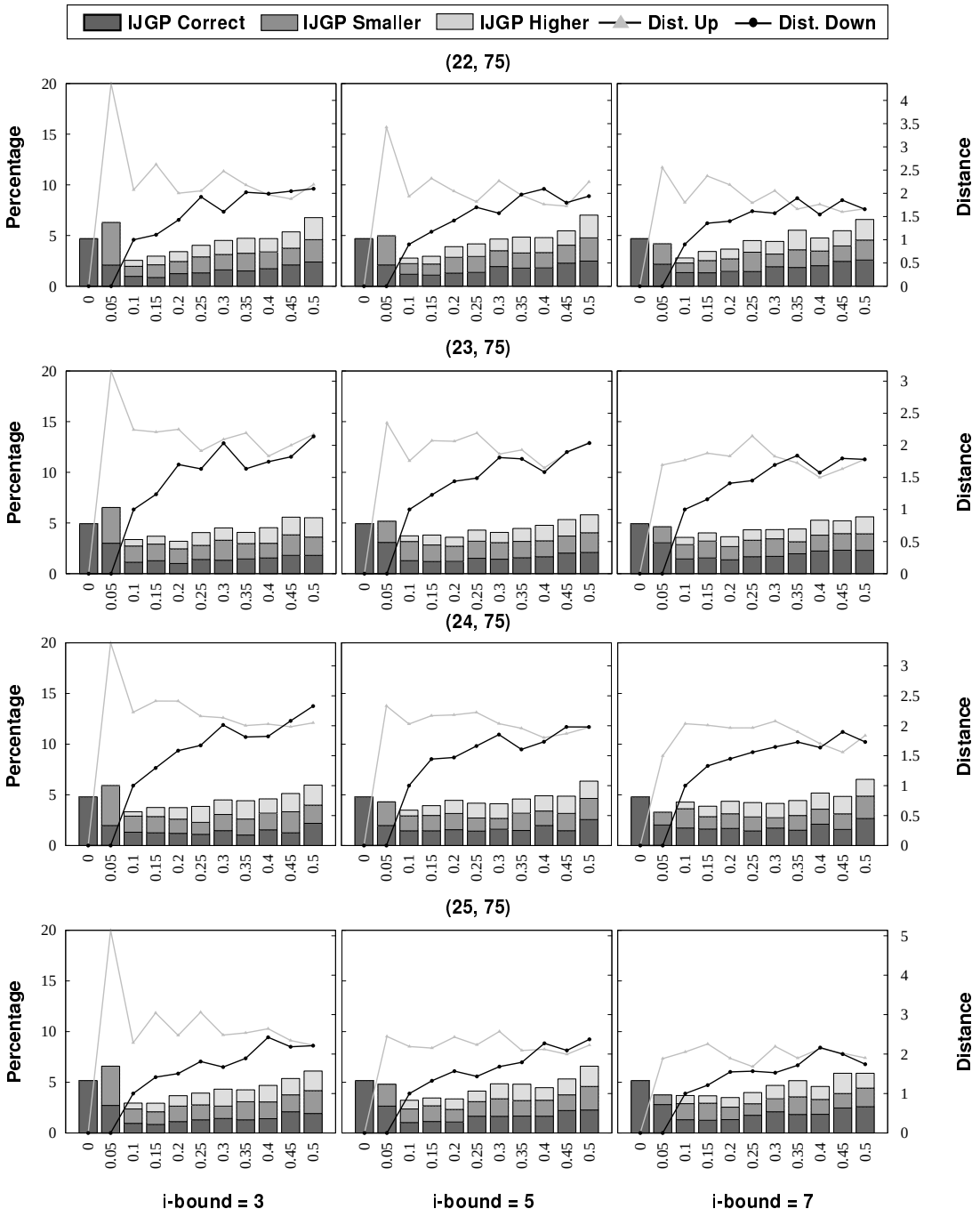

Figure 20: *Distance* results on grids instances. $\epsilon = 0.05$

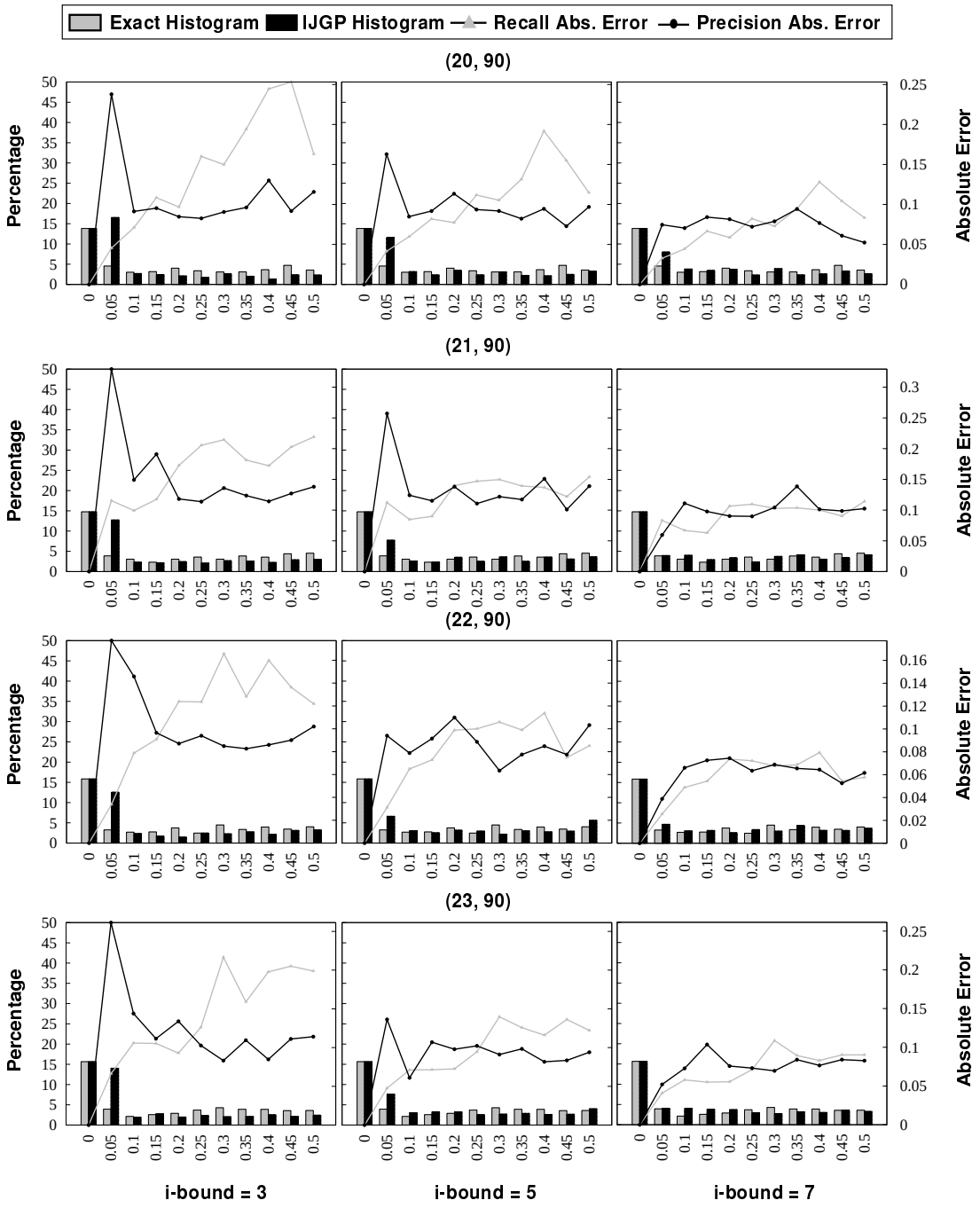

Figure 21: *Absolute error* results on grids instances. Each row shows the results for one parameter configuration (N, D). Each parameter configuration has $N \times N$ variables and one evidence node. The induced width w* is 29, 31, 32 and 33, respectively. $\epsilon = 0.05$

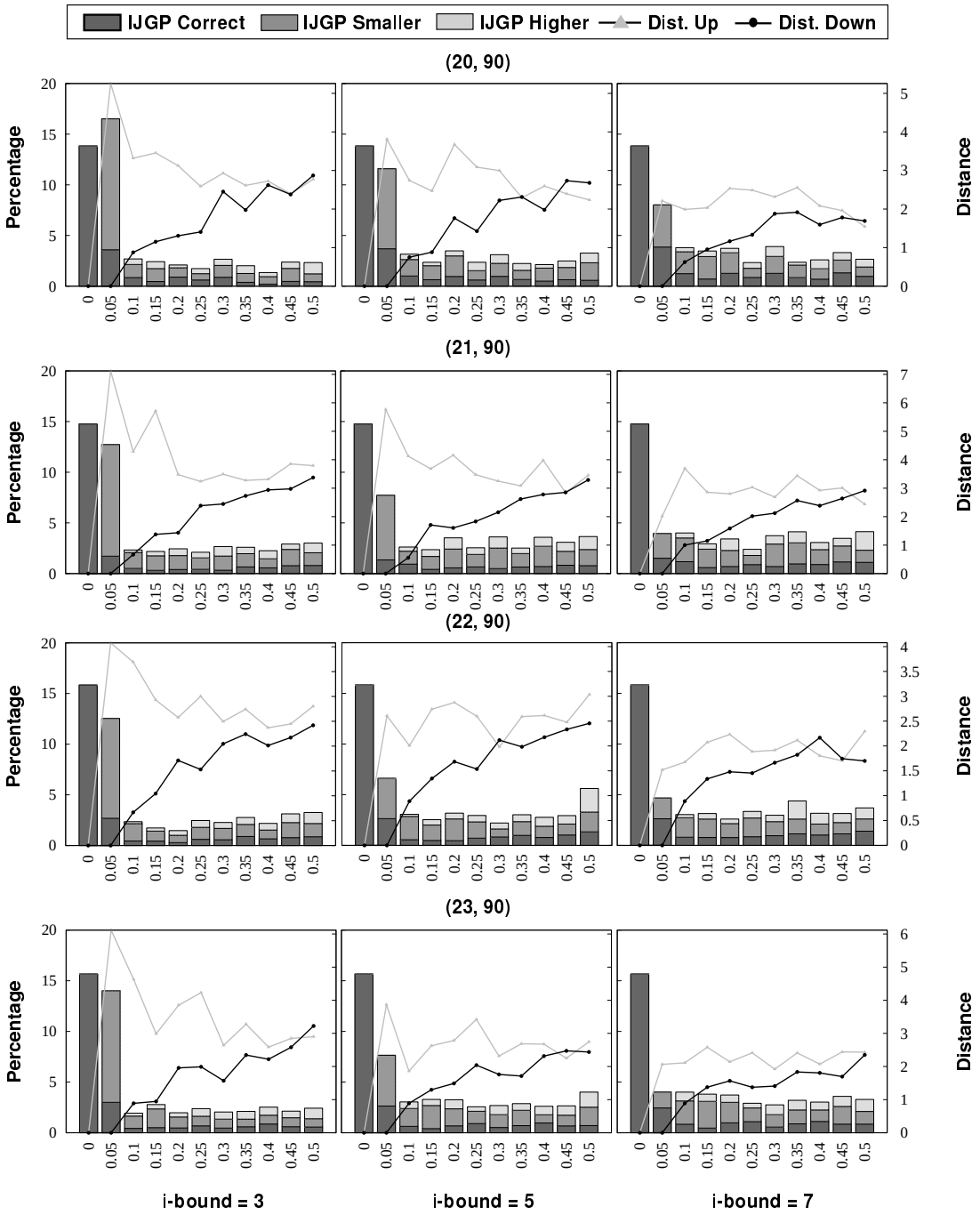

Figure 22: *Distance* results on grids instances. $\epsilon = 0.05$

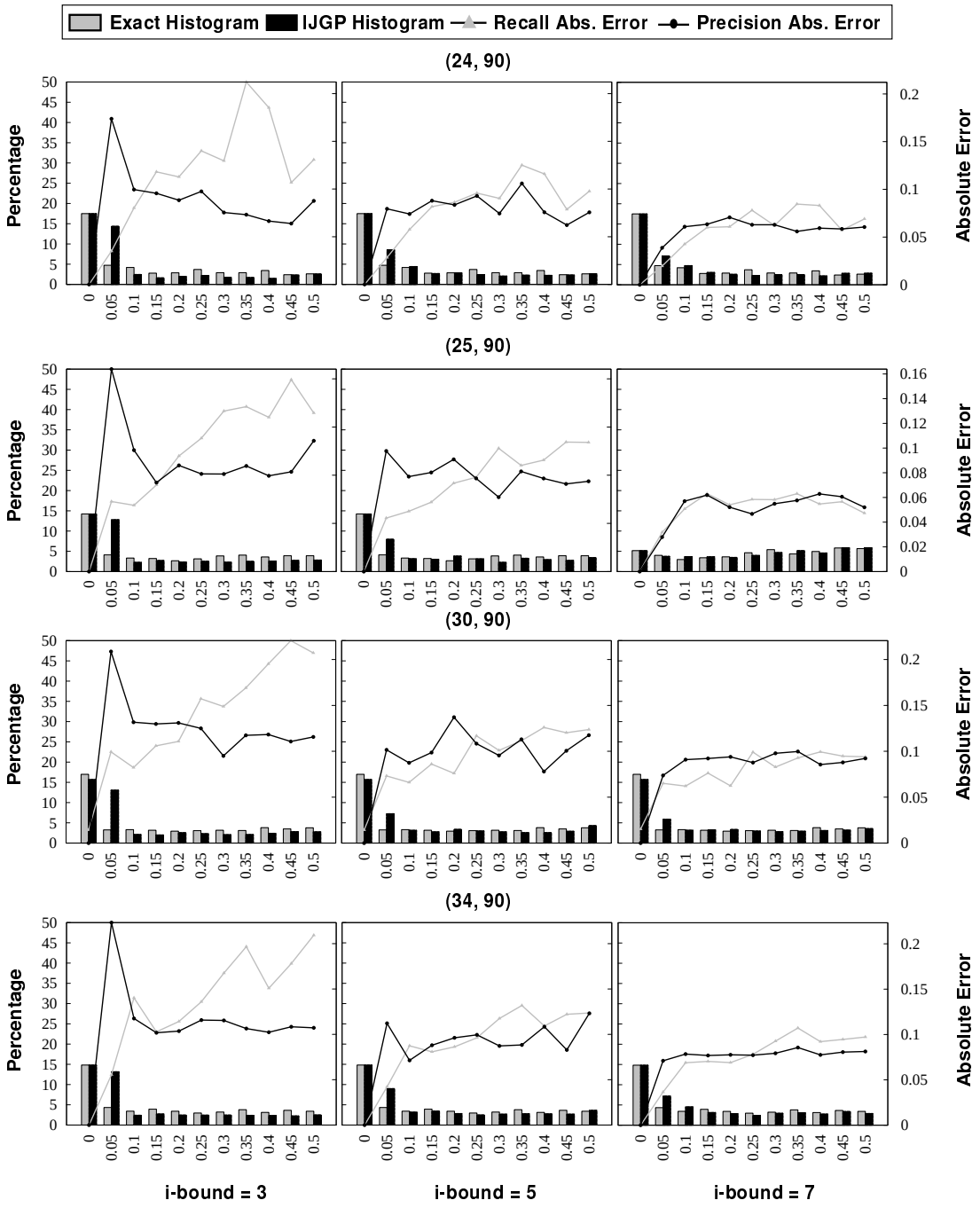

Figure 23: *Absolute error* results on grids instances. Each row shows the results for one parameter configuration (N, D). Each parameter configuration has $N \times N$ variables and one evidence node. The induced width w* is 36, 38, 46 and 53, respectively. $\epsilon = 0.05$

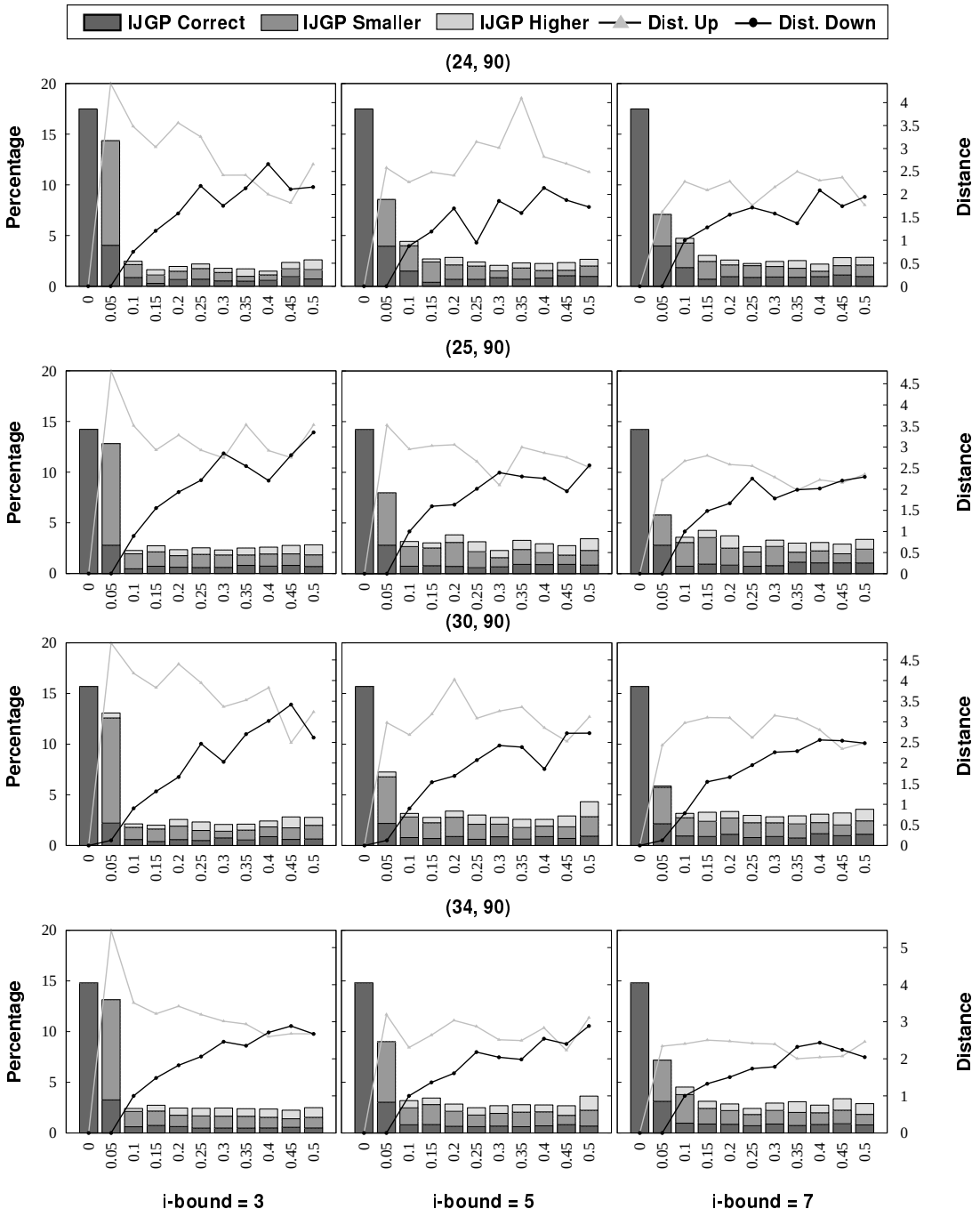

Figure 24: *Distance* results on grids instances. $\epsilon = 0.05$

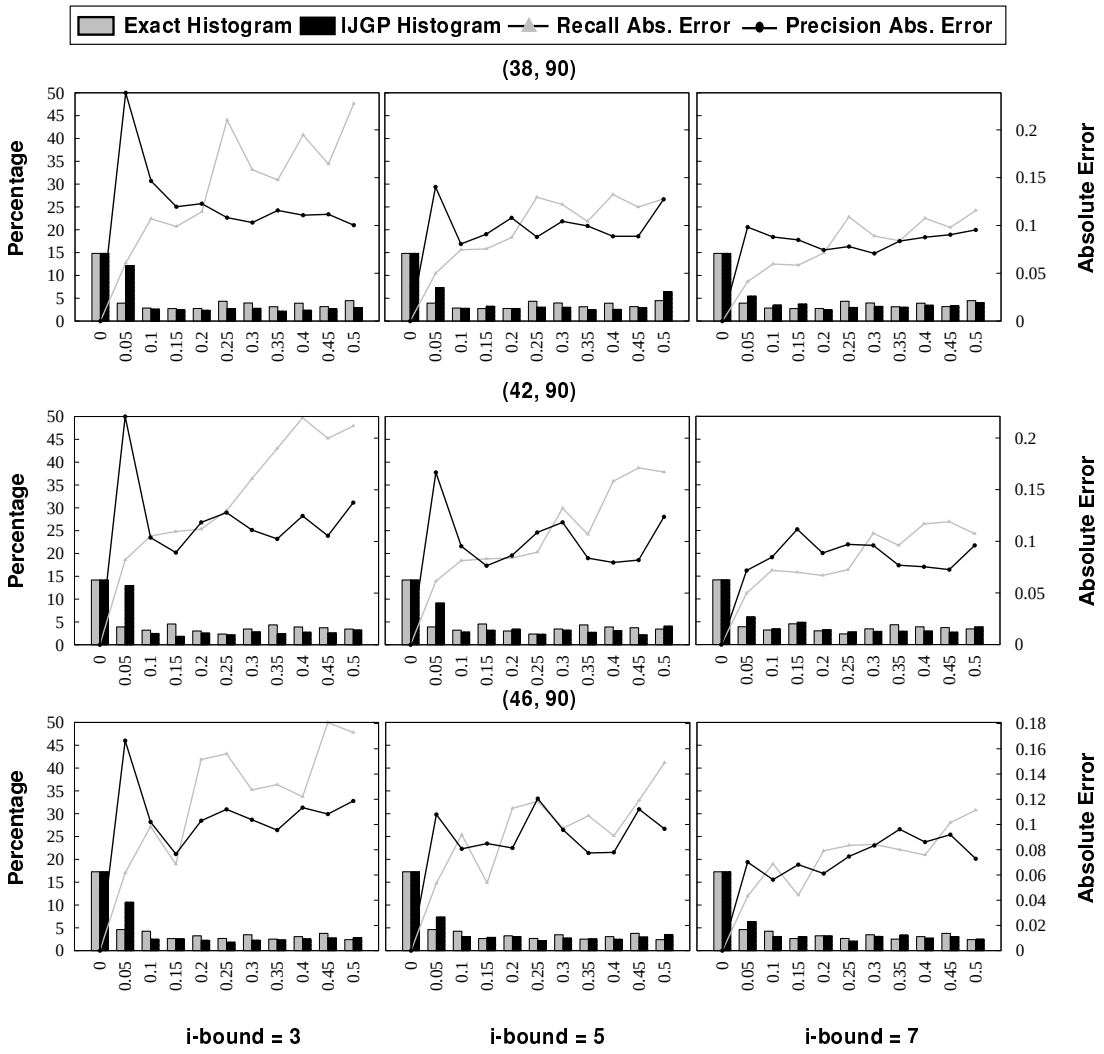

Figure 25: *Absolute error* results on grids instances. Each row shows the results for one parameter configuration (N, D). Each parameter configuration has $N \times N$ variables and one evidence node. The induced width w* is 61, 70 and 77, respectively. $\epsilon = 0.05$

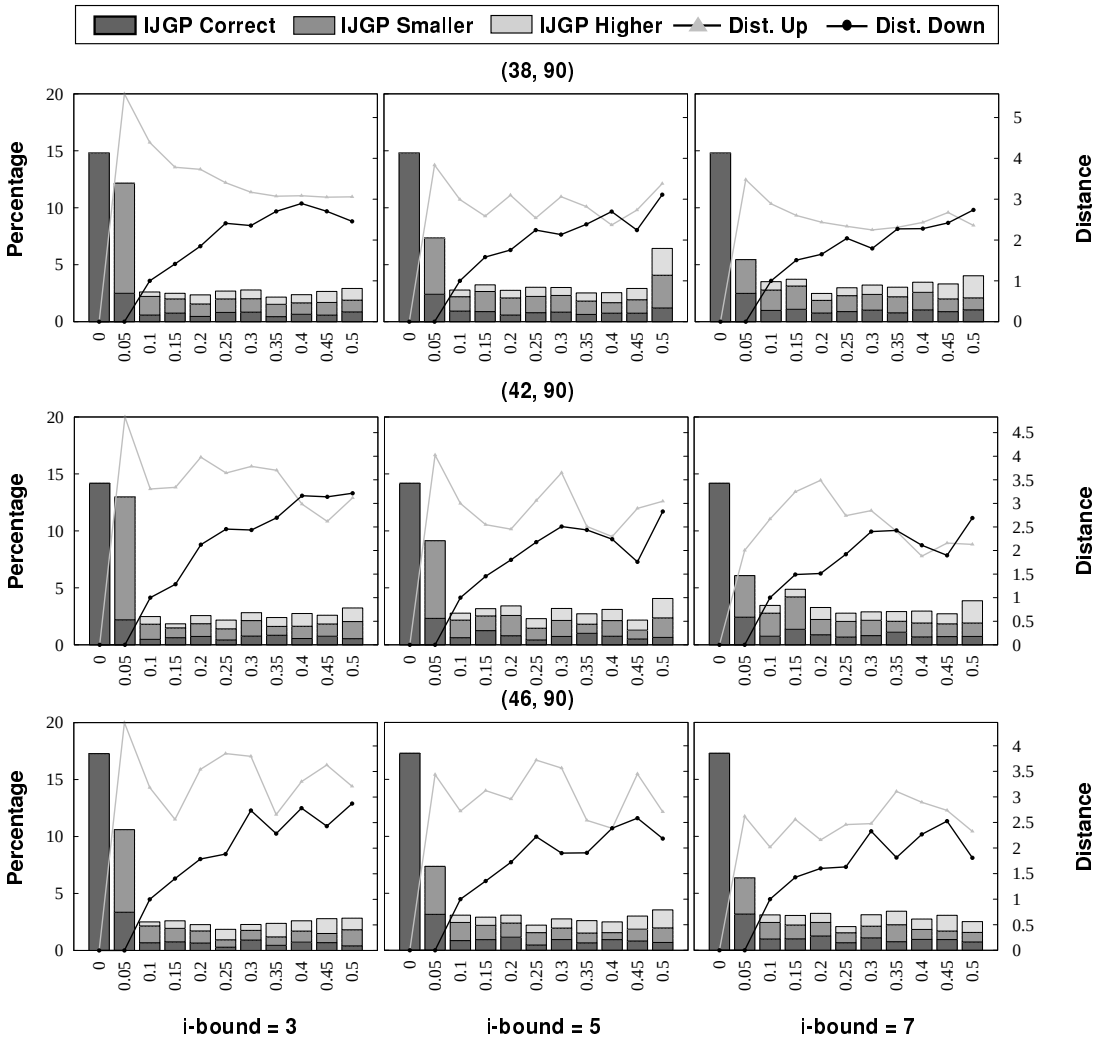

Figure 26: *Distance* results on grids instances. $\epsilon = 0.05$

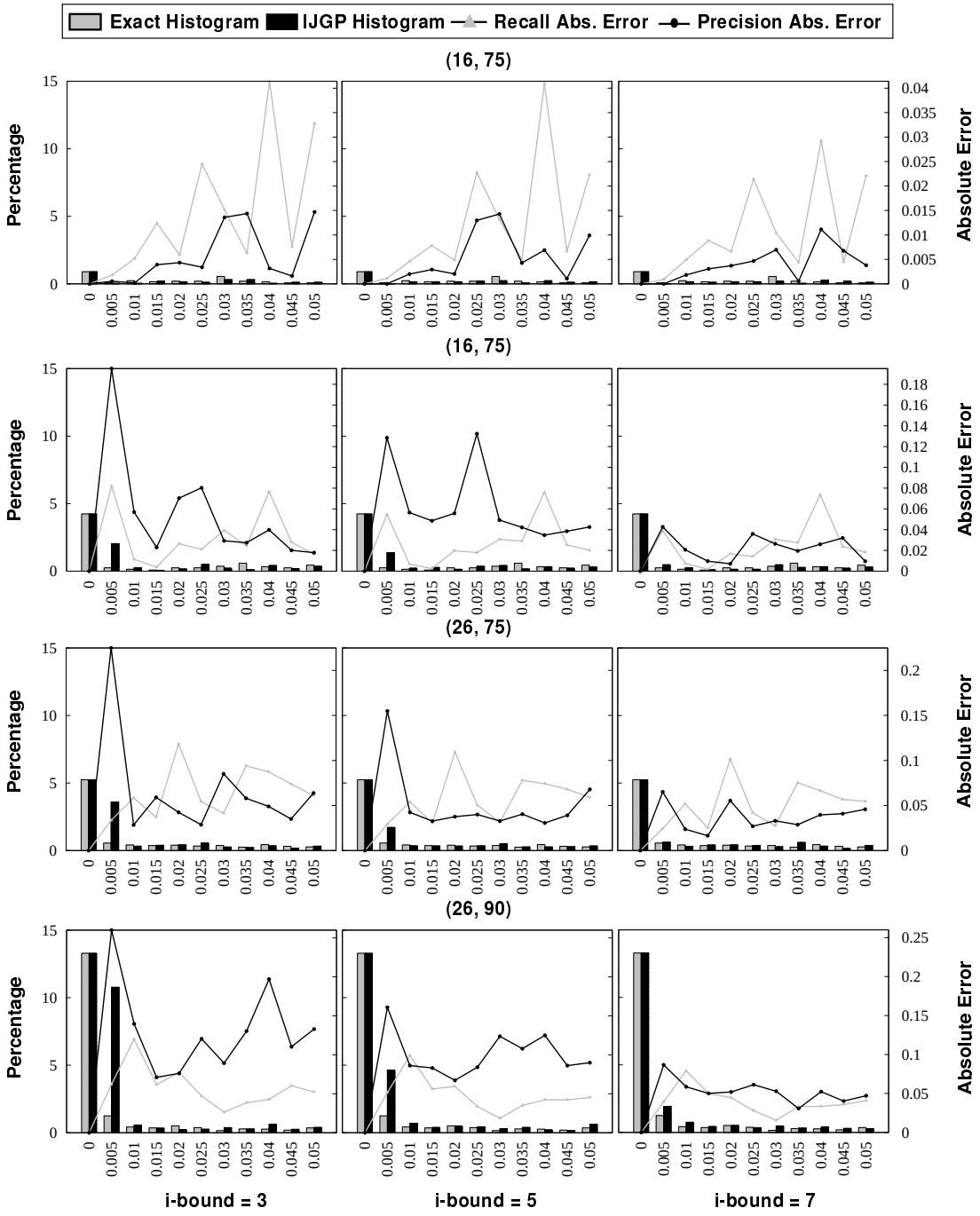

Figure 27: *Absolute error* results on grids instances. $\epsilon = 0.005$.

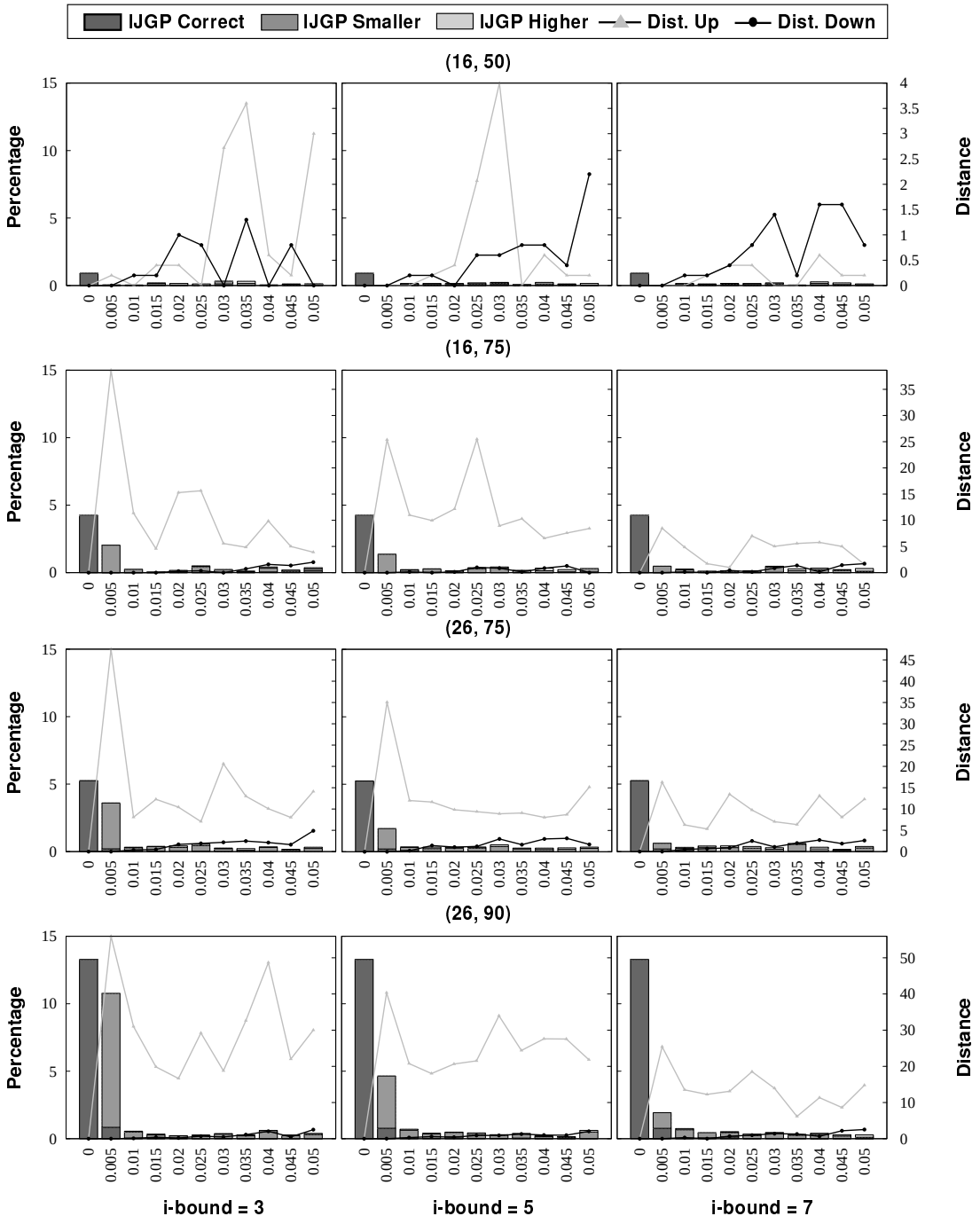

Figure 28: *Distance* results on grids instances. $\epsilon = 0.005$.

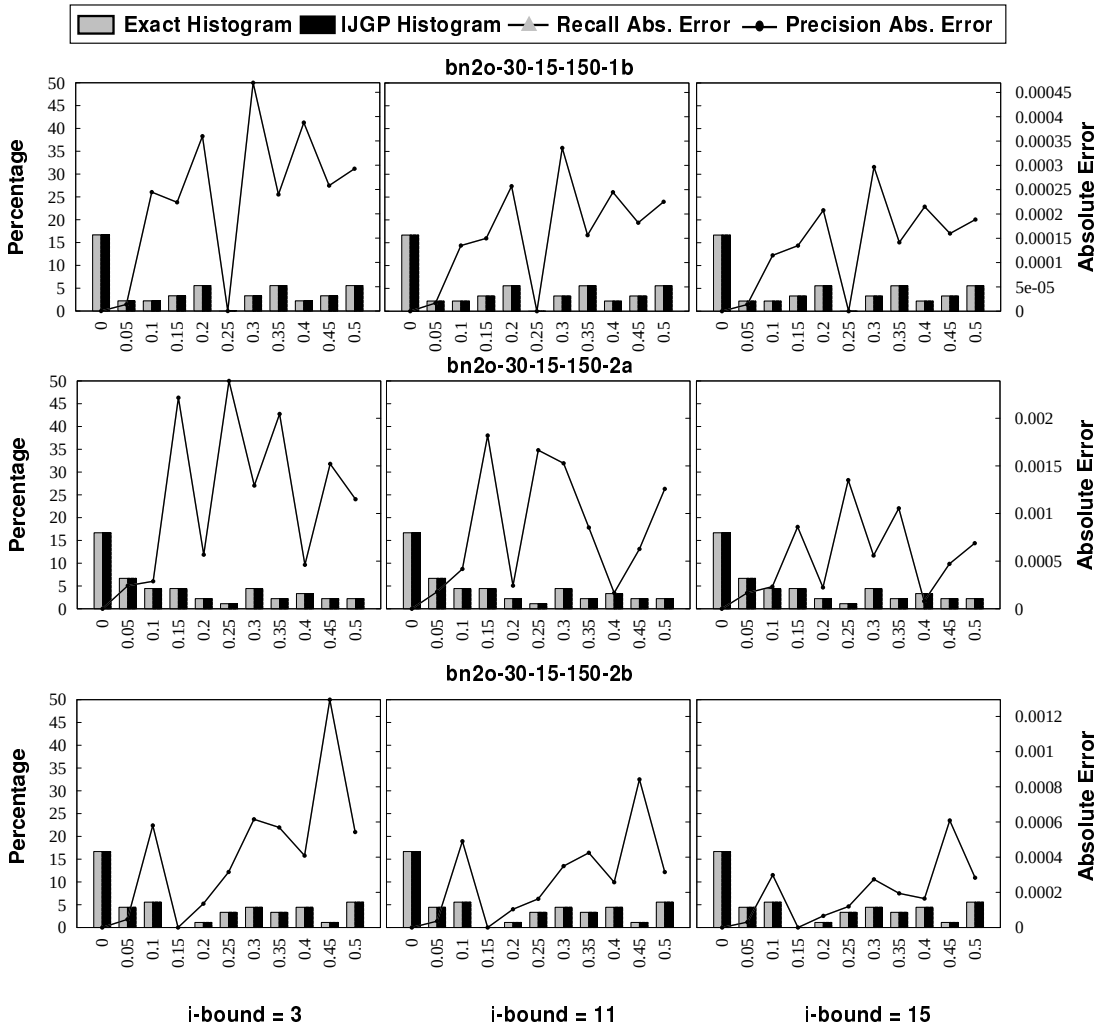

Figure 29: *Absolute error* results on bn2o instances. The number of variables N, number of evidence variables NE, and induced width w* of each instance is $N = 45$, $NE = 15$, and w*=24. $\epsilon = 0.05$

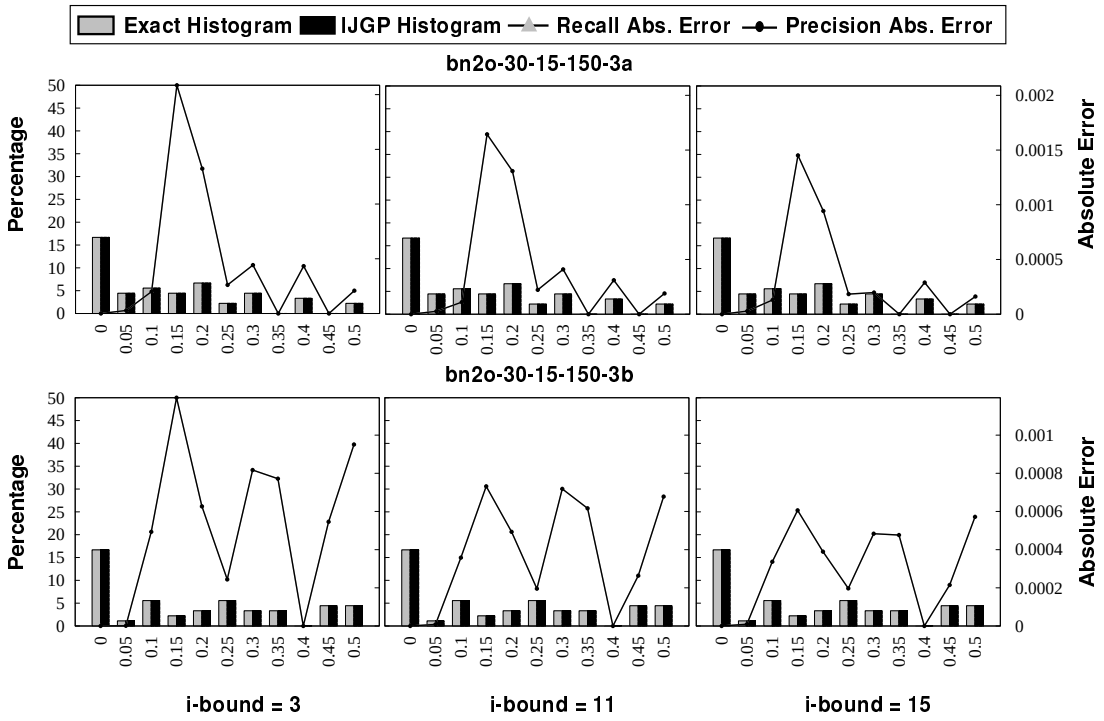

Figure 30: *Absolute error* results on bn2o instances. The number of variables N, number of evidence variables NE, and induced width w* of each instance is $N = 45$, $NE = 15$, and w*=24. $\epsilon = 0.05$

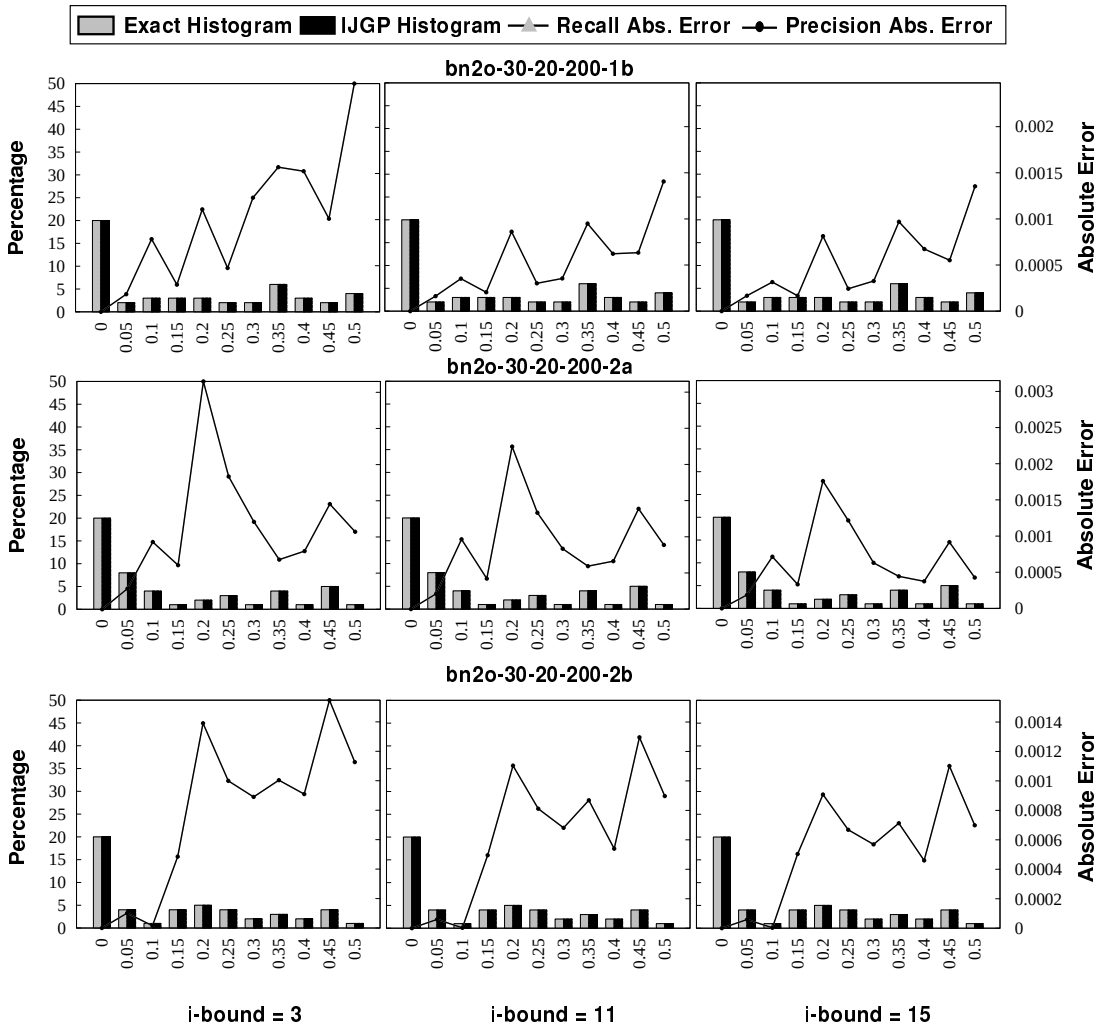

Figure 31: *Absolute error* results on bn2o instances. The number of variables N, number of evidence variables NE, and induced width w* of each instance is $N = 50$, $NE = 20$, and w*=27. $\epsilon = 0.05$

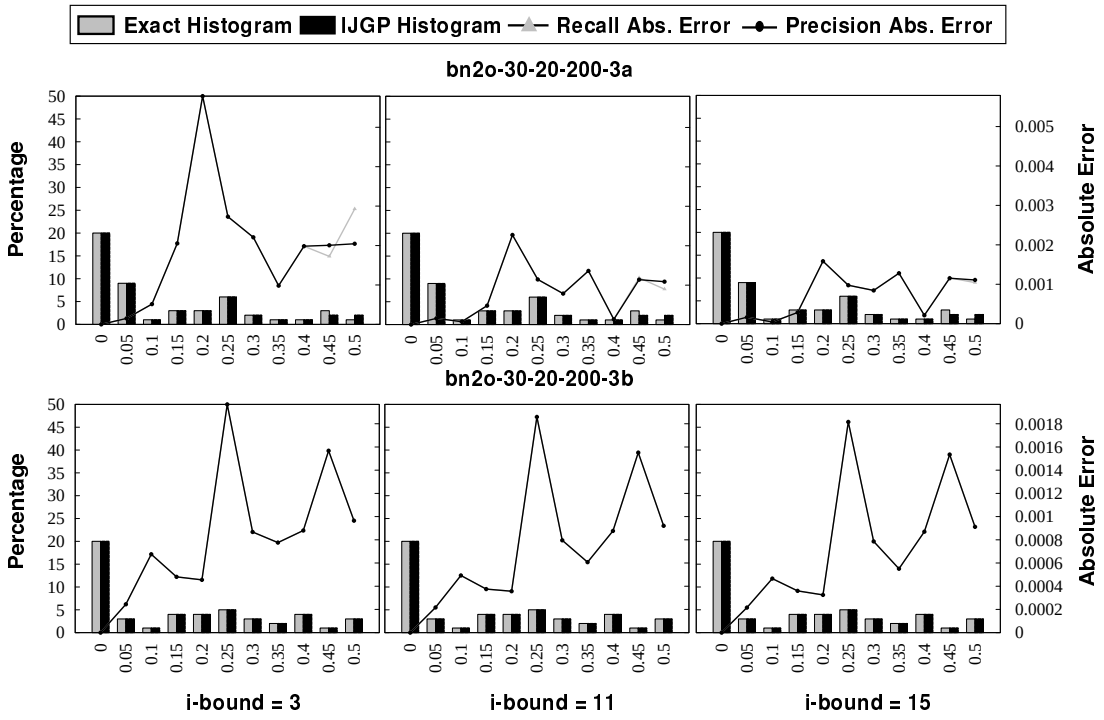

Figure 32: *Absolute error* results on bn2o instances. The number of variables N, number of evidence variables NE, and induced width w* of each instance is $N = 50$, $NE = 20$, and w*=27. $\epsilon = 0.05$

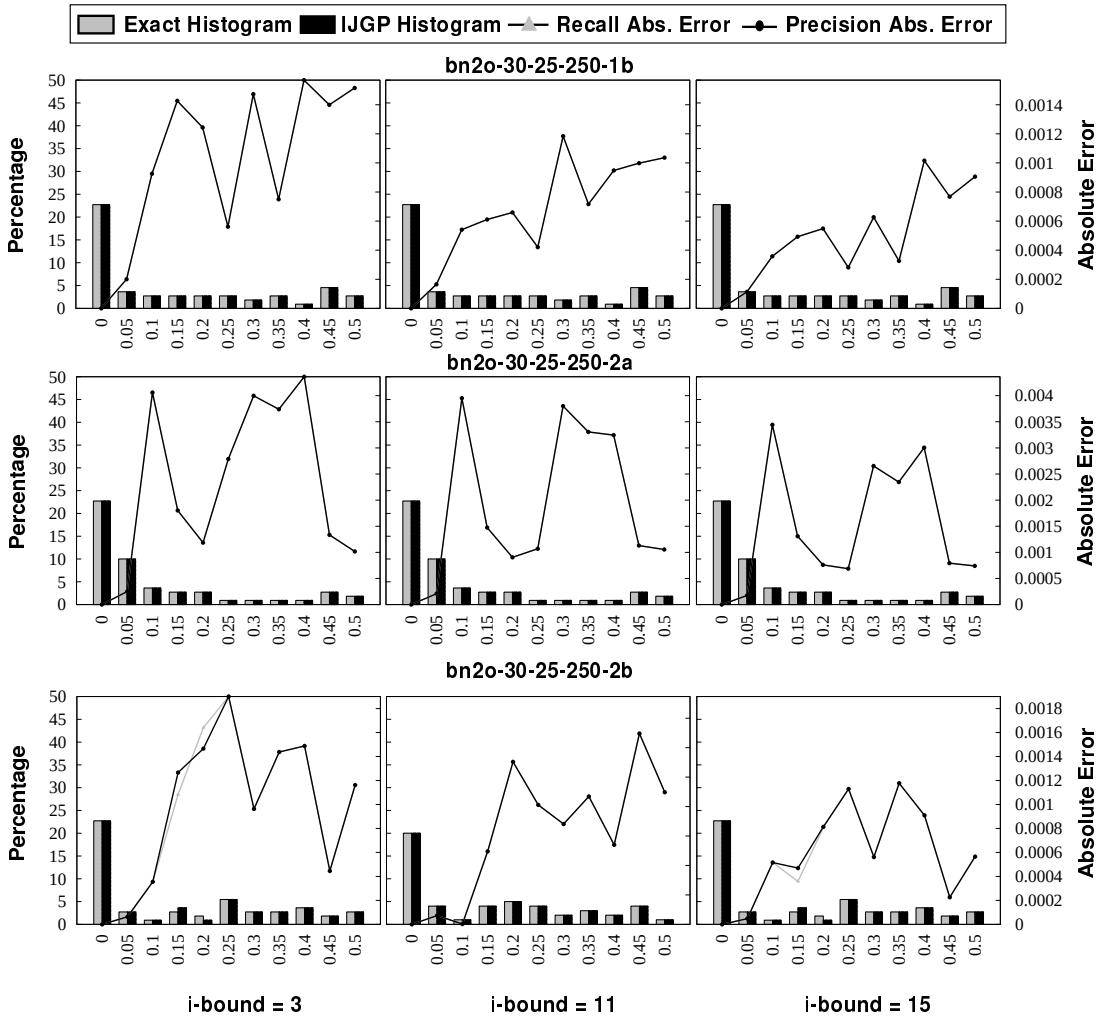

Figure 33: *Absolute error* results on bn2o instances. The number of variables N, number of evidence variables NE, and induced width w* of each instance is $N = 55$, $NE = 25$, and w*=26. $\epsilon = 0.05$

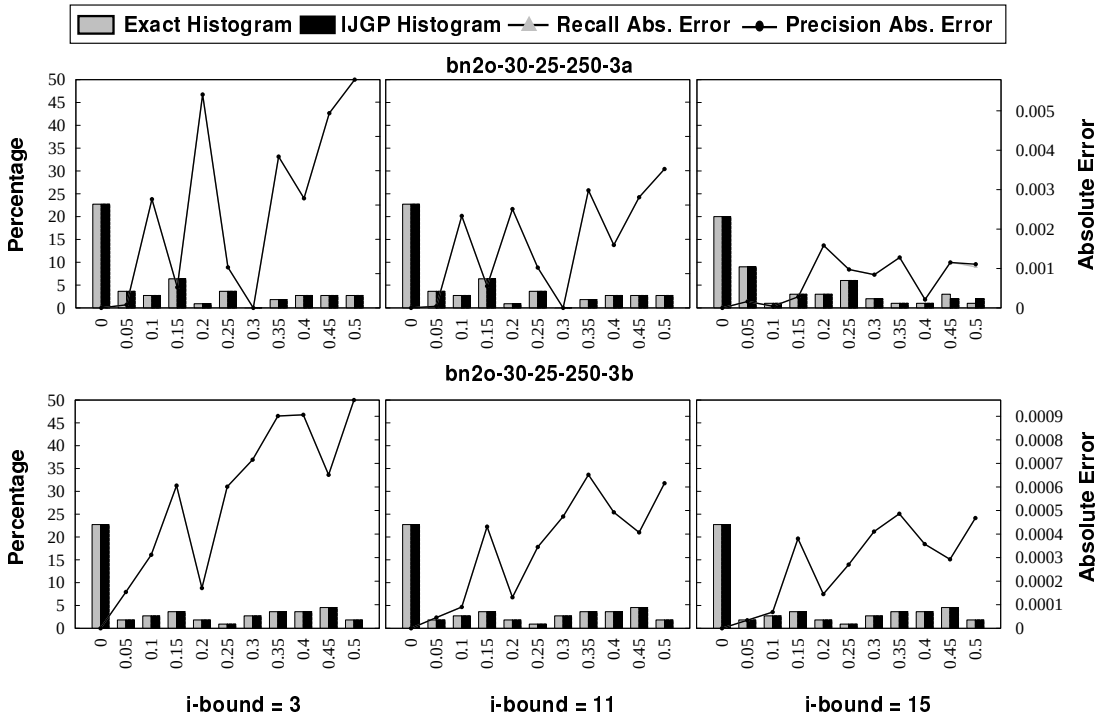

Figure 34: *Absolute error* results on bn2o instances. The number of variables N, number of evidence variables NE, and induced width w* of each instance is $N = 55$, $NE = 25$, and w*=26. $\epsilon = 0.05$

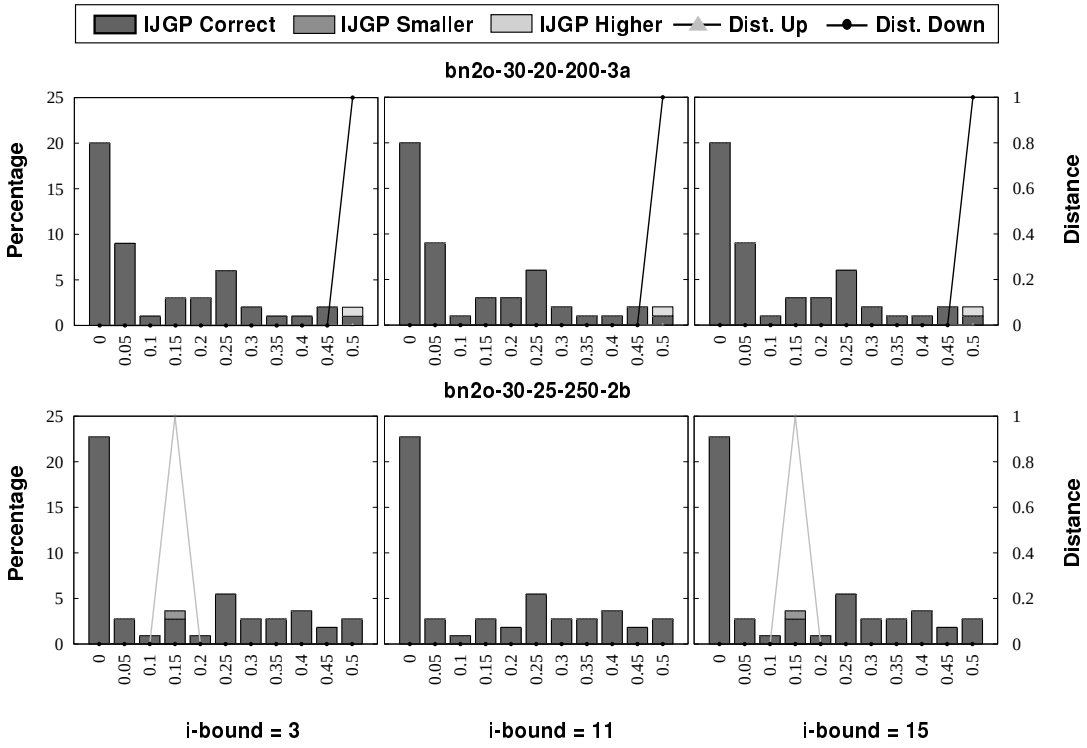

Figure 35: *Distance* results on bn2o instances. $\epsilon = 0.05$