The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

Analysis of Region-Based Integral Calculus Problems

Chris Alvin
Furman University
chris.alvin@furman.edu

Adam Byerly
Bradley University
abyerly@fsmail.bradley.edu

Abstract

Students in Integral Calculus courses solve a canonical class of problems based on regions. We present a technique for identifying and dissecting such regions into integrable components. We show that our techniques result in a minimal set of constituent regions and demonstrate the utility of our techniques experimentally by solving "area between curves" problems, classifying each problem with a relative difficulty rank, and comparing our ranking to the ranking implied by textbooks.

1 Introduction

In a traditional *Introduction to Integral Calculus* course, students solve "area between curves" (*problems* when context is clear). Figure 1 depicts such an area problem consisting of functions adapted from (Stewart 2007). Our solving and analysis consists of (1) identifying the regions defined by the problem, (2) solving the area problem (i.e. setting up a definite integral expression that computes the area of the defined region(s)) with respect to (w.r.t.) a vertical axis of integration, (3) dissect each region into sub-regions for solving w.r.t. a horizontal axis of integration, and (4) computing difficulty-based features.

Our first step identifies the exact set of regions defined by the functions and domain of the problem. We do so by computing the facets of a planar graph corresponding to the functions in the area problem. In Figure 1, the two functions f and h define a single region $\boldsymbol{r}$. For each region, we intuitively label the boundaries: left, right, top, and bottom. For example, region $\boldsymbol{r}$ in Figure 1 consists of a left boundary point $(-1, 0)$, right boundary point $\left(\frac{1}{2}, \frac{3}{4}\right)$, top boundary function $h(x) = -x^2 + 1$, and bottom boundary function $f(x) = x^2 + x$. Given this 'rectangular' view of region $\boldsymbol{r}$ in Figure 1, it is clear that the area can be computed w.r.t. the x-axis with solution integral bounds $x = -1$ and $x = \frac{1}{2}$. Since our region is defined by a single top and bottom function, we may construct the solution integral w.r.t. x as $\int_{-1}^{1/2} \left[\left(-x^2 + 1\right) - \left(x^2 + x\right)\right] dx = 9/8$.

As a feature for describing problem difficulty, we solve all such area problems w.r.t. both the x-axis and the y-axis. With respect to the x-axis, both f and h in Figure 1 are not one-to-one over the interval $x \in \left[-1, \frac{1}{2}\right]$; hence, region $\boldsymbol{r}$ must be dissected into sub-regions that are integrable w.r.t. the $y-$axis. Consider solving w.r.t. y by sliding a horizontal line over $\boldsymbol{r}$ from the maximum of $h(x)$ down to minimum of $f(x)$. This action demonstrates that there is no single top / bottom function when integrating w.r.t. y. Specifically, the dashed line segment labeled ~~dy~~ (dy with strikethrough) demonstrates that $f(x)$ cannot be both a top and bottom function in an integral solution in $y \in [-1/4, 0]$; similarly, $y \in [3/4, 1]$ prohibits $h(x)$ as a top and bottom function. The dotted segment dy in Figure 1 identifies distinct top and bottom functions for $y \in [0, 3/4]$. Finding the area of region $\boldsymbol{r}$ in Figure 1 w.r.t. the y-axis requires summing the areas of the five constituent sub-regions $r_1, \ldots, r_5$. We first compute f^{-1} and h^{-1} and select $+$ or $-$ as necessary to treat f^{-1} and h^{-1} as functions. Our solution for the area of r is thus given by $\int r_1 \, dy + \ldots + \int r_5 \, dy$.

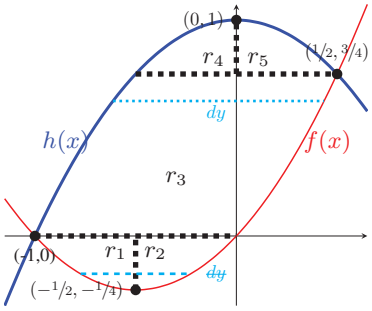

Figure 1: $f(x) = x^2 + x$ and $h(x) = -x^2 + 1$ Resulting in Region $\boldsymbol{r} = \cup_{i=1}^{5} r_i$

Our first contribution dissects a region into a minimal number of subregions for solving w.r.t. y. Second, we posit that there is a correlation between the perception of the difficulty of a problem and the number of dissected regions required to find the area of the region w.r.t. the y-axis.

2 Preliminaries

Integrable regions. Our definition of closed, integrable regions is based on integrable functions forming a topologically equivalent rectangular structure. A *region* is a Jordan curve (Jordan 1893) embedded in the Euclidean plane. Thus,

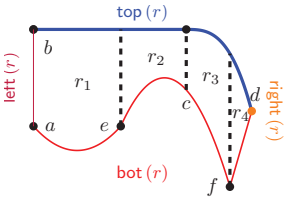

Figure 2: A Piecewise Defined Region

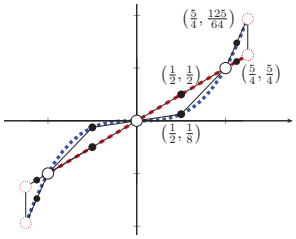

Figure 3: Planar Graph for $p = \langle \{x, x^3\}, [-1.25, 1.25] \rangle$

there exists a homeomorphism between a region r and a rectangle embedded in the Cartesian Plane. We can topologically equate our notion of a region and the simple notion of a rectangle by labeling the sides as left, right, top, and bottom as shown in Figure 2.

Let z be the independent variable for a set of functions defining a region in the Cartesian Plane. A *left (resp. right) bound* for a region r is either a line segment defined in terms of z or a single point. A *top (resp. bottom) bound* for a region r is an integrable function $h(z)$ (resp. $f(z)$). An *integrable region* r consists of a left bound, right bound, top bound, and bottom bound, all defined w.r.t. z.

In Figure 2 we observe $\mathsf{left}\,(r)$ is a vertical segment from point a to b while $\mathsf{right}\,(r)$ consists of point d; we acquire the horizontal (x) component of the left (resp. right) region bound using $x_{\mathsf{left}(r)}$ (resp. $x_{\mathsf{right}(r)}$). The top bound is a piecewise defined function: a horizontal segment in the interval $[x_b, x_c]$ and a cubic polynomial $[x_c, x_d]$. The bottom bounds consist of two quadratic polynomials in the intervals $[x_a, x_e]$ and $[x_e, x_f]$ followed by a linear polynomial in the interval $[x_f, x_d]$.

We refine our notion of an integrable region to account for integrating w.r.t. y. We say a *singleton region* is an integrable region with both top and bottom bounds defined by integrable functions that are not piecewise-defined. Region $\boldsymbol{r}$ in Figure 1 is a singleton region. Further, region r_3 defined w.r.t. x in Figure 1 is not a singleton region; however, region r_3 defined w.r.t. y is a singleton region. Last, all subregions, save r_3 in Figure 1, are dual singleton (singleton regions w.r.t. x and y).

Problems and solutions. We say a *region-based problem* is a tuple of the form $p = \langle F, D \rangle$ where F is a set of integrable functions and D a domain. If the domain is unspecified ($D = \emptyset$) then p requires the student compute an *implicit domain*. For $p = \langle \{x^2 + x, -x^2 + 1\}, \emptyset \rangle$ in Figure 1, solving $x^2 + x = -x^2 + 1$ results in the implicit domain $\left[-1, \frac{1}{2}\right]$. In comparison, $\langle \{x^2 + x = -x^2 + 1\}, [-2, 0] \rangle$ has two distinct regions w.r.t. x.

An *area between curves problem* is a region-based problem p with the goal of computing the area of all constituent regions defined by p. With respect to problem difficulty, we say problems $p_1 \equiv p_2$ have equivalent difficulty ($\mathsf{difficulty}\,(p_1) = \mathsf{difficulty}\,(p_2)$) and similarly, $p_1 \prec p_2$ when $\mathsf{difficulty}\,(p_1) \leq \mathsf{difficulty}\,(p_2)$.

Definition 1 (Area Between Curves Solution)**.** *Let p be an area between curves problem $\langle F, D \rangle$ that defines a set of n regions $\{r_i\}$. The solution to p is given by* $\mathsf{A} = \sum_{i=1}^{n} \mathsf{Area}\,(r_i)$. *For a region s partitioned into m intervals $I_1, \ldots, I_m$ defining singleton subregions $s_1, \ldots, s_m$, the area of s is given by* $\mathsf{Area}\,(s) = \sum_{j=1}^{m} \mathsf{Area}\,(s_j)$. *Then* $\int_{x_{\mathsf{left}(\ell)}}^{x_{\mathsf{right}(\ell)}} [\mathsf{top}\,(\ell) - \mathsf{bot}\,(\ell)]\, dx$ *defines the area of a singleton region ℓ w.r.t. x. Similarly,* $\int_{y_{\mathsf{bot}(\ell)}}^{y_{\mathsf{top}(\ell)}} [\mathsf{right}\,(\ell) - \mathsf{left}\,(\ell)]\, dy$ *for ℓ defined w.r.t. y.*

If problem $p = \langle F, D \rangle$ defines a non-singleton region r, we must partition r into a sequence of singleton sub-regions R. According to Definition 1, the solution to an area problem is a sequence of integral expressions. The area of region r in Figure 2 w.r.t. x consists of four constituent singleton regions $\{r_1, r_2, r_3, r_4\}$ defined by the partition $x_a < x_e < x_c < x_f < x_d$: $\mathsf{Area}\,(r) = \int_{x_a}^{x_e} [\mathsf{top}\,(r_1) - \mathsf{bot}\,(r_1)] + \int_{x_e}^{x_c} [\mathsf{top}\,(r_2) - \mathsf{bot}\,(r_2)] + \int_{x_c}^{x_f} [\mathsf{top}\,(r_3) - \mathsf{bot}\,(r_3)] + \int_{x_f}^{x_d} [\mathsf{top}\,(r_4) - \mathsf{bot}\,(r_4)]\, dx$.

3 Region Analysis

We describe algorithms for dissecting a region into a cover (Munkres 2013) of disjoint sub-regions. We assume an implicit, finite domain and range; without this assumption, $p = \langle \{\sin x, 0\}, \emptyset \rangle$ defines an infinite number of regions.

Identifying regions. To capture all regions defined by a set of functions F over domain D in a problem $p = \langle F, D \rangle$, we take two steps: (1) construct a planar graph w corresponding to p and (2) identify the facets in w. The facets we identify correspond directly to the regions defined by p. We detail planar graph construction and defer to (Alvin et al. 2017) for facet identification.

Without loss of generality we describe construction of a planar graph w defined in terms of x. For a problem $p = \langle F, D \rangle$, the first step is to identify all points of intersection among all functions $f \in F$ in domain D: $I_f = \cup_{f_1, f_2 \in F} \{(x, y) \mid f_1 \neq f_2 \wedge x \in D \wedge f_1(x) = f_2(x)\}$. All points in I_f are added to w. We observe x and x^3 in Figure 3 intersect at $x = -1, 0, 1$ in the domain $[-1.25, 1.25]$.

For well-constructed problem $p = \langle F, D \rangle$ with an implicit domain, we assume finite bounds defined by the functions in F. With explicit domain D, if the endpoints are not intersection points (e.g., Figure 3), we construct vertical segments at the endpoints of D. For each respective bound at x, we add an edge (x, m) to (x, M) to w where $m = \min\{y \mid f(x), \forall f \in F\}$ and $M = \max\{y \mid f(x), \forall f \in F\}$ representing a vertical segment. We add two such edges to the planar graph at $x = \pm 1.25$ in Figure 3.

In its current state, w consists of the set of points from endpoints of the domain and intersection points (e.g., open circles in Figure 3). Ambiguity arises if we connect these points with edges in w. Hence, for all $f \in F$, we add to w the 'midpoints' between all points in $I_f \cup \{(x, m), (x, M)\}$ that lie on all $f \in F$. w is thus a planar graph that corresponds unambiguously to the functions in F. In Figure 3, our disambiguating points are the smaller, solid points and the solid lines make up the planar graph w.

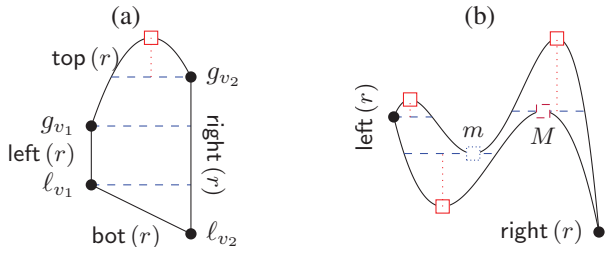

Figure 4: Dissecting a Singleton Region r w.r.t. the y-axis

Computing singleton regions. Computing the area of a region requires identifying two distinct bounds: a 'top' function and a 'bottom' function. For a non-singleton region, we partition the region into singleton sub-regions. For region r defined in terms of x, we merge the partition endpoints of the top and bottom of r. In Figure 2, the ordered endpoints are $E = \{x_a, x_e, x_c, x_f, x_d\}$. For all non-extreme points in E, we construct vertical line segments. The result is two outer 'bookend' regions with at least one bound being a vertical line. The left 'bookend' (r_1 in Figure 2) leverages the existing left bound of r, a new vertical segment, and the top and bottom functions defined in the interval. Similarly, for the right 'bookend' (r_4 in Figure 2). For inner regions such as r_2 and r_3, we iteratively construct all singleton regions with two vertical segments.

Region dissection with respect to y. For a *singleton* region r defined w.r.t. x, our algorithm computes a minimal set of disjoint sub-regions S_r, each defined w.r.t. y such that $r = \cup_{s \in S_r} \{s\}$. For a function $y = f(x)$, $f^{-1}(y)$ may be a relation; integrable regions w.r.t. the y-axis requires we segment the relation into functions over one-to-one intervals.

For a left (resp. right) bound as a vertical line segment, we construct a horizontal segment *into* r, if possible. That is, we consider the least lexicographic point (ℓ) and add a horizontal segment if the bottom bound is decreasing w.r.t. x from ℓ. For example, in Figure 4(a), the bottom linear bound of r, $\mathsf{bot}(r)$, is decreasing at the point ℓ_{v_1}; therefore, we extend a (dashed) horizontal segment into r. Similarly, we add a horizontal segment from a point g if the top bound is increasing at g: e.g., the top bound in Figure 4(a) at g_{v_1}.

When the left bound is a point, we insert a horizontal segment from $\mathsf{left}(r)$ into r that does not extend beyond the first intersection with the top bound. No segment is added from $\mathsf{right}(r)$ in Figure 4(b) since $\mathsf{bot}(r)$ and $\mathsf{top}(r)$ are both decreasing at $\mathsf{right}(r)$.

We further decompose a function $f(x)$ into one-to-one intervals by constructing horizontal segments from each minimum of the top function (resp. maxima of the bottom) in the given domain. In Figure 4(b), a segment is added from local minimum m and local maximum M as those points are both in the interval $\left[x_{\mathsf{left}(r)}, x_{\mathsf{right}(r)}\right]$.

Adding only horizontal segments in Figure 1 creates three sub-regions with one being integrable w.r.t. y (indicated with dy); the other two sub-regions are problematic because each have the same 'left' and 'right' bounds w.r.t. y (indicated with ~~dy~~). By Rolle's Theorem (Stewart 2007), since our functions are differentiable, there exists an extreme point between the endpoints. We argue informally that the bound is a horizontal line or there is only one such extreme point (otherwise, we would have constructed more horizontals from minima). Thus, singleton regions w.r.t. y are created by extending vertical segments from the maxima in the domain of the top and similarly for the minima of the bottom bound in its domain (as guaranteed by Rolle's Theorem). For example, Figure 1 constructs two vertical segments resulting in five total sub-regions; three are constructed in Figure 4(b).

Our algorithm decomposes a singleton region r defined w.r.t. x into a minimal number of sub-regions. Furthermore, each sub-region is a singleton region w.r.t. y. A formal proof is omitted due to space limitations.

Table 1: Problem Features from Least-Significant (1) to Most-Significant (23)

1	Complex Graphing	**13**	Restrictive Domain
2	Has Trigonometric	**14**	Has Polynomials
3	Method of How Solution Bounds are Computed	**15**	More Easily Solved by Opposing Axis
4	No. Problem Regions	**16**	No. Functions Defined by x
5	'Other' Function Present	**17**	No. Dissected Subregions
6	Has Exponential	**18**	Domain Stated
7	Rational Function	**19**	Other Variables Used (z)
8	Sum of Polynomial Function Degrees	**20**	No. of Function Bounds
9	Largest Degree Polynomial	**21**	Has Logarithms
10	No. Functions Defined by y	**22**	Has Root Functions
11	Regions are Symmetric	**23**	Has Piecewise Function
12	Degree of Root Function		

4 Experimental Analyses

We selected 97 area problems from two seminal Calculus textbooks (Stewart 2007; Larson and Hostetler 1986) and identified a set of features that, in our opinion, would contribute to the perceived difficulty of a problem.

Background. We propose a formal model of relative problem difficulty we call *step-wise difficulty*. This model is a well-ordering of problems based on their difficulty where each pair of problems $p_{2n+1} \equiv p_{2n+2}$ for $n \in \mathbb{N}_0$ and $p_{2n+1} \prec p_{2n+3}$ for $n \in \mathbb{N}_0$.

Problem Features. Table 1 orders all of the features used to build our models from least significant (1) to most significant (23). To identify the feature orders, we built a linear regression model that initially included 23 features as potentially correlative with problem difficulty. We then used the backward elimination form of stepwise regression (Derk-

sen and Keselman 1992): we repeatedly removed the feature from the model with the least statistical significance and then rerunning the model until the only features remaining in the model showed a statistical significance (p-value < 0.05). For example, the method for computing the integral bounds is experimentally critical in determining problem difficulty. On the contrary, few problems in the corpus contain logarithms due to difficult integrability, thus it is a negligible feature based on our sample. Sadly, our analyses by y described in §3 were not the most salient features (15 and 17).

Survey of Educators. We surveyed 26 high school and college educators asking "What are some of the most important factors in determining the difficulty of an 'Area Between Curves' problem?" *Graphing.* Region identification is contingent on properly graphing the problem functions and is one of the most significant features (Table 1): 62% of survey respondents identified graphing as a factor. *Points of Intersection.* Solving the system of equations may be simple (i.e., factoring a quadratic), using more complex algebra (i.e., a trigonometric function and polynomial), or requires a calculation device to compute the real-valued irrationals. Since basic algebra can be a problem for students, this issue was noted by 65% of the experts. *Multiple Regions.* The more regions in a problem, the more complex the integral expression solution. 27% of the experts made direct or indirect reference to this idea and according to Table 1 is one of the most salient features for determining problem difficulty. *Variables.* 65% of the experts said students would find the problem functions $\{x^2 - 6x, 0\}$ less difficult than $\{z^2 - 6z, 0\}$ with 25% saying they were equivalent. We accounted for variables as a feature, but it ranked 13$^{\text{th}}$ in significance. *Calculus.* Only 27% of experts mentioned Calculus-related issues (antidifferentitation and evaluation) as contributing to the difficulty of a problem. Our feature set did not take into account complexity of integration.

Feature analysis and difficulty model. We constructed a multi-layer perceptron (MLP) to test the ability of the identified features to predict the relative difficulty of the problems. Our hypothesis was that the order of the problems in a textbook is indicative of problem difficulty as described by the identified features. The input to the MLP consisted of the sets of the features from pairs of problems with the goal of predicting which problem comes earlier in a set of textbook problems. The MLP used two hidden layers (resp. 32 and 8 neurons), the Adam optimizer (Kingma and Ba 2014) with default parameters, ReLU activation, a maximum of 1000 epochs, and early convergence termination.

We constructed two sets of inputs for the MLP. The first being a *naive* pairing of all problems in each textbook section: 1124 pairs. The second being the same pairing, excluding adjacent pairs when the first problem was an odd numbered problem: 1078 pairs. We ran each MLP 100 times with random train/test splits with 15% of the samples being reserved for testing. For each test set accuracy, we calculated the mean and standard deviation discarding runs with results more than two standard deviations from the mean.

Our ground truth for problem difficulty is textbook ordering. We conducted baseline naive and baseline stepwise experiments by excluding region-specific features (4, 11, 15, 17 in Table 1); results are shown in Table 2. We compared our baseline accuracies to the complete feature set and found statistically significant improvements by incorporating the region-based features: $p < 0.03$ t-tests for both naive and stepwise. The experiments (Table 2) show that both the naive and stepwise versions are strong predictors of problem difficulty. We also performed t-tests of the two experimental results for each trial and achieved an average p-value of 0.0143, indicating stepwise is a statistically significantly better predictor of problem difficulty. We repeated the experiments using logistic regression with random train/test splits of 50%, rather than a MLP. These experiments agreed that the stepwise version was a statistically significantly better predictor: mean accuracies of the logistic regression model were 0.77324 and 0.77893 for the naive and stepwise variants, respectively. We conclude that MLP has greater accuracy than a logistic model in predicting relative textbook problem difficulty, suggesting a non-linear relationship among the predictor features.

Table 2: Accuracy of Naive and Step-Wise MLP Models (Averaged Over 10 Runs)

	Min	Mean	Max	Std. Dev.
Baseline Naive	0.776	0.828	0.880	0.028
Naive	0.785	0.835	0.885	0.027
Baseline Step-Wise	0.793	0.843	0.893	0.027
Step-Wise	0.801	0.850	0.901	0.027

5 Conclusions

We have described algorithms for identifying regions and solving area between curves problems, a steadfast application in Integral Calculus. Our region analyses identified a feature set descriptive of area between curves problems and their relative difficulty. We then demonstrated the utility of our algorithms by developing a statistically significant ($p < 0.03$) non-linear model showing such problems are ordered in textbooks according to difficulty and step-wise difficulty.

References

Alvin, C.; Gulwani, S.; Majumdar, R.; and Mukhopadhyay, S. 2017. Synthesis of solutions for shaded area geometry problems. In *FLAIRS*, 14–19.

Derksen, S., and Keselman, H. J. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* 45(2):265–282.

Jordan, C. 1893. *Cours D'analyse de L'École Polytechnique*. Gauthier-Villars et fils.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

Larson, R., and Hostetler, R. 1986. *Calculus*. D.C. Heath and Company, third edition.

Munkres, J. R. 2013. *Topology*. Pearson.

Stewart, J. 2007. *Calculus*. Available 2010 Titles Enhanced Web Assign Series. Cengage Learning.