1. Implement Gradient Descent

A general-purpose gradient descent library was created which provides for specification of x_0 (the initial guess), η (the step size), and ϵ (the convergence threshold: if two consecutive steps differ by less than this value, the algorithm terminates). The gradient descent procedure was tested on two functions with well-known optimal values: (i) a non-convex polynomial $f(x)$, and (ii) a negative bivariate Gaussian $p(x)$:

$$f(x) = x^4 - x^3 - x^2$$

$$p(\mathbf{x}) = \frac{-100}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Where in $p(\mathbf{x})$, $\boldsymbol{\Sigma}$ represents the covariance and $\boldsymbol{\mu}$ the mean. Note also that the standard Gaussian has been multiplied by -1 (so that we can pose our optimization as a minimization) and scaled by 100, which is just to help with making plotting more clear.

The effects of initial guesses (x_0) and varying step size (η) are best explained via the plots. For the non-convex polynomial, $f(x)$, the effect of the starting position can be seen from the three plots in Figure 1. Initial guesses f$x_0 > 0$ will terminate at the global maximum of the function, whereas initial guesses $x_0 < 0$ get "stuck" at the local minimum in $\mathbb{R}^-$. An initial guess of exactly a local maximum $x_0 = 0$ leads to the algorithm terminating still with $x_final = 0$ after only a handful of iterations, since initially $f'(x=0) = 0$. It is nice to note that local maxima can cause algorithm termination with poor initial guesses, but in practice adding randomization to initial guesses helps mitigate this problem. The effect of the step size can be viewed by observing the counter plots for gradient descent on $p(\mathbf{x})$, the bivariate Gaussian. In plots (a) and (b), the step size is sufficiently small and successful descent occurs to the global minimum. In plot (c), however, the step size is sufficiently large such that the algorithm continually jumps "back and forth" over the minimum, and the algorithm terminates only due to the maximum number of function calls ($n_{max} = 5000$ was used).

Numerical gradients were also implemented by calculating finite differences. Note that each gradient descent update accordingly requires two function evaluations (per dimension...). A comparison of results for analytical vs. numerical gradients is provided in the three tables below. For all points tested, analytical and numerical gradients resulted in the same termination value, although analytical gradients were typically able to find this value in approximately 1/5th the function calls.

SAY SOMETHING ABOUT CONVERGENCE CRITERION

Number of function calls

Function and Initial Guess, x_0	Analytical	Numerical	Scipy
Non-convex $f(x)$			
$x_0 = 1.9$	130	646	24
$x_0 = -1.0$	312	1556	36
$x_0 = 0.0$	2	6	3
Neg. Gaussian $p(x)$			
$x_0 = (1.0, 1.0)$	177	1603	48

Minimum x

Function and Initial Guess, x_0	Analytical	Numerical	Scipy
Non-convex $f(x)$			
$x_0 = 1.9$	1.17	1.17	-0.43
$x_0 = -1.0$	-0.43	-0.43	1.17
$x_0 = 0.0$	0.0	$2.5e-13$	0
Neg. Gaussian $p(x)$			
$x_0 = (1.0, 1.0)$	$(0, 0.5)$	$(0, 0.5)$	$(0, 0.5)$

Minimum function value

Function and Initial Guess, x_0	Analytical	Numerical	Scipy
Non-convex $f(x)$			
$x_0 = 1.9$	-1.10	-1.10	-0.07
$x_0 = -1.0$	-0.07	-0.07	-1.10
$x_0 = 0.0$	0.0	$-6.3e-26$	0
Neg. Gaussian $p(x)$			
$x_0 = (1.0, 1.0)$	-17.36	-17.36	-17.36

2. Linear Basis Function Regression

We implemented the linear basis function regression in python and were able to get good agreement of our plots and regression weights with those in Bishop (MAYBE INCLUDE SOME PLOTS HERE???). The closed form solution to the linear least squares problem is convenient, however it requires inverting the matrix $\Phi^T\Phi$. An alternative to avoid this matrix inversion is to apply gradient descent to the sum of squared error (SSE) objective function given by $(\Phi w - y)^T(\Phi w - y)$. Differentiating the SSE with respect to w gives a gradient of $2\Phi^T(\Phi^T w - y)$. Using this analytical gradient we can apply our gradient descent code from problem 1. Our code employs a fixed step size of η and the termination criterion is as soon as $\epsilon_{n+1} = |SSE(w_n) - SSE(w_{n+1})| \leq \gamma$ where γ is a tolerance parameter. Our initial parameter choices are $\eta = 0.05, \gamma = 1 \times 10^{-8}$. The initial guess is set to $w_0 = -w_{OLS}$ where w_{OLS} are the true regression weights.

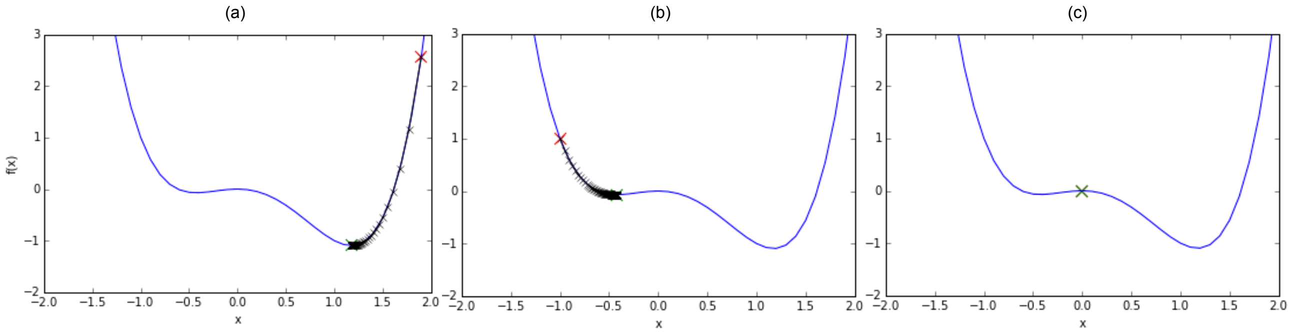

Figure 1. Visualization of gradient descent for three initial guesses, x_0, of the non-convex polynomial $f(x) = x^4 - x^3 - x^2$. Initial guesses are: (a), 1.9; (b) -1.0; (c), 0.0. The red X marks the initial guess, the green X marks the algorithm's final value, and the black X represent sequential values produced by each gradient descent step. Plot (a) converges to the global minimum, while (b) converges to a local, non-global minimum. Plot (c) remains at the initial guess, since $f'(x_0) = 0$. Shown are results using: analytical gradients $\eta = 0.02$, $\epsilon = 0.0004$.

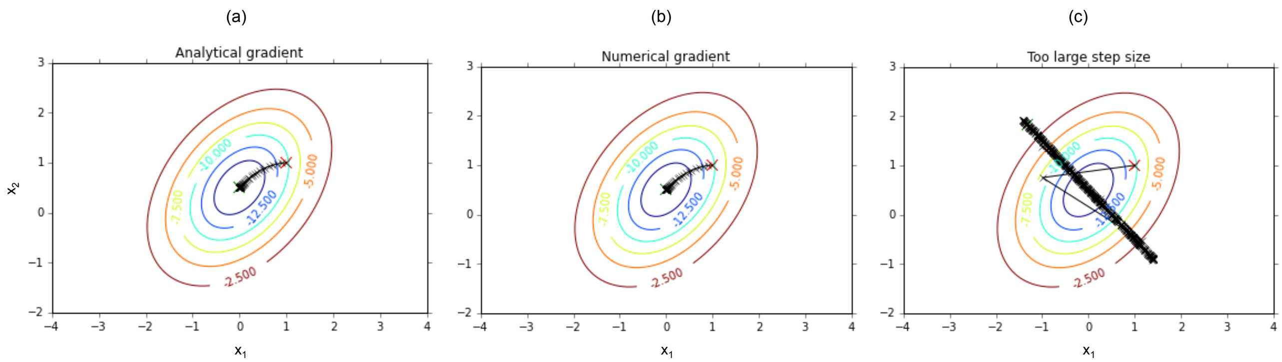

Figure 2. Visualization of gradient descent for the negative bivariate Gaussian $p(x)$, plotted via contours. The red X marks the initial guess, the green X marks the algorithm's final value, and the black X represent sequential values produced by each gradient descent step. Plots (a) and (b) show a comparison of using the analytical vs. numerical gradients: note although the descent path looks similar, the number of function calls is an order of magnitude different. Plot (c) shows the path of gradient descent when the step size is too large ($\eta = 0.2$). Unless otherwise noted, all plots used: $x_0 = (1.0, 1.0)$, $\eta = 0.01$, $\epsilon = 0.0004$.

For $M = 1$ gradient descent works quite well.

Solver	Function Calls	Weights
Gradient Descent	113	$(0.820, -1.267)$
Scipy	20	$(0.820, -1.267)$

Our gradient descent method converges to essentially the same regression weights as the scipy.optimize.minimize method, albeit with an order of magnitude more function calls. This is to be expected as the SSE function is convex in w and hence there is a unique global minimum which can be reached by "following" the gradient downhill. Changing the initial guess to all zeros produces very similar performance in terms of objective value and number of function calls.e With $M = 3$ however the performance of our solver changes dramatically however. In particular, if w_{OLS} denote the true regression weights then

Solver	Function Calls	$\lvert w - w_{OLS} \rvert_2$	η
Grad. Desc.	$38,513$	0.14	0.05
Grad. Desc.	$72,657$	0.199	0.025
Scipy	114	4×10^{-5}	-

Thus increasing the dimension of the optimization problem from 2 to 4 exposes the weaknesses of our gradient descent method. In particular it uses about 300 more function calls. The reason is that as we approach the optimum the SSE function becomes very flat and hence the gradient ∇SSE becomes very small. Hence, in our update step $w_{n+1} = w_n - \eta * \nabla SSE(w_n)$ the amount we move, $\eta * \nabla SSE(w_n)$ is becoming arbitrarily small. In particular if we plot function value vs iterations (Figure 3) we see that the plot is becomes very flat quite quicky.

This is a result of our update step not moving us enough when the gradient becomes small. Reducing η to 0.025 just

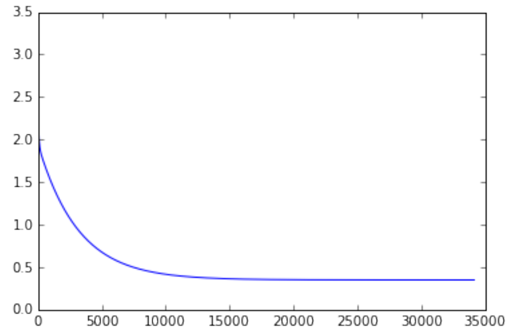

Figure 3. Value of SSE at each iteration of our gradient descent algorithm, $\eta = 0.05$, $w_0 = 0$

exacerbates this problem. In this case we require more iterations and achieve worse fit of the regression weights. If we try increasing $\eta = 0.07$ however the opposite happens. We adjust w_{n+1} by too much and jumping to the other side of the "bowl" representing the SSE. Hence we end up bouncing back between different sides of the bowl and the solution ends up exploding. The effect of the inital guess isn't too important. Using $\eta = 0.05$ and choosing initial guess to the origin results in the number of function calls decreasing to about $34,000$ and doesn't change the achieved accuracy of the solution. This makes sense because since we are optimizing a convex function we have no risk of getting stuck in local minima and thus independent of where we start we should be able to follow the gradient down to the minimum.

The results for $M = 9$ are very interesting as well. The value of $SSE(w_{OLS})$ is $1e-7$. If we set the initial guess for the scipy minimization to be $w_0 = 1.5 * w_{OLS}$ then it achieves a minimum value of 0.09 however if we set $w_0 = 0$ then the method only achieves $SSE = 0.314$. Our own gradient descent method behaves similarly. Thus the performance of the gradient descent methods depends heavily on the initial guess. I believe that this is because since we are considering large powers of x the function is very flat near the minimum. Thus gradient descent methods have a hard time making progress since locally the function is almost flat.

If the basis functions were of the form $\phi_n(x) = \sin(2\pi nx)$ then we would expect a regression vector of the form $w = (1, 0, 0, ..., 0)$ since the data was actually generated from $\sin(2\pi x)$ with Gaussian noise added. A potential disadvantage is that since the sin function is periodic you are imposing periodicity on your data. In particular x and $x+1$ will map to the same value.

3. Ridge Regression

As we saw in the previous section $M = 3$ provides a fairly good fit to the data. As can be seen in Figure 4 using $M = 3, \lambda = 0.01$ doesn't provide good performance.

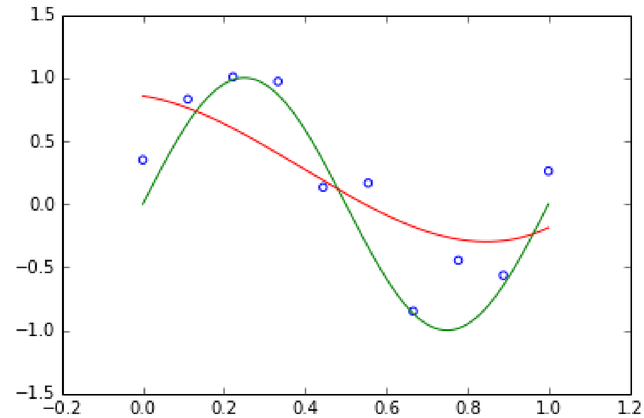

Figure 4. Ridge regression with $M = 3, \lambda = 0.01$. $SSE = 1.54$

In particular the λ weight is too high and keeps the weights close to zero, which results in a relatively flat straight line. Reducing λ to 0.001 as in Figure 5 produces better results, as can be seen by comparing the SSE's.

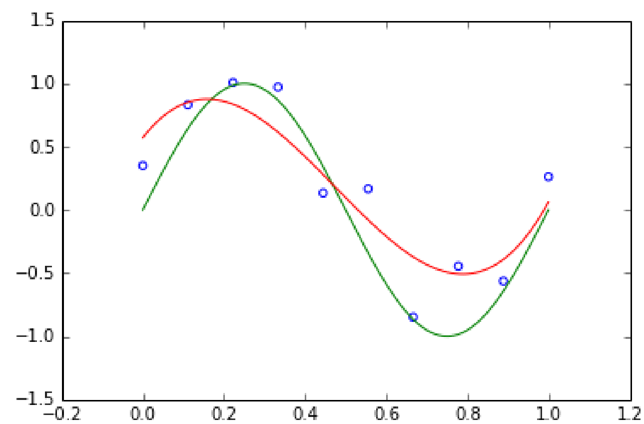

Figure 5. Ridge regression with $M = 3, \lambda = 0.001$. $SSE = 0.588$

The interesting thing is that increasing M while keeping λ fixed doesn't adversely affect performance. In particular the prediction curve in red looks fairly similar even as we increase the number of features to $M = 9$ in Figure 6.

Thus the regularizer λ helps us to avoid overfitting, contrary to the standard OLS regression as in the previous section.

For the train, validate and test datasets we perform the following model selection procedure. Given (M, λ) we run ridge regression on the training dataset to find the regression weights $w(M, \lambda)$. Then we compute the SSE using weights $w(M, \lambda)$ on the validation dataset, denote this by $SSE_v(M, \lambda)$. We choose (M, λ) to minimize

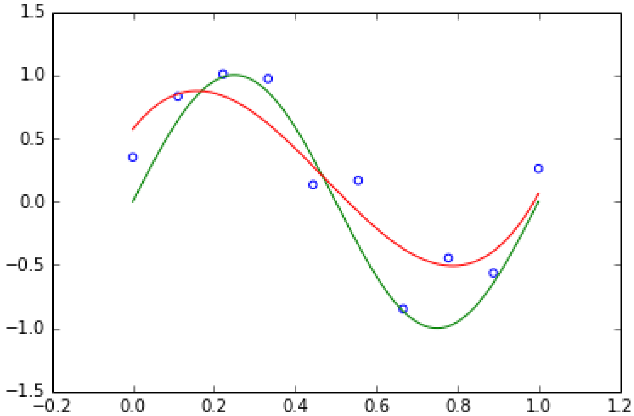

Figure 6. Ridge regression with $M = 9, \lambda = 0.001$. $SSE = 0.418$

M	λ	SSE Train	SSE Validate	SSE Test
1	6.53	18.36	16.89	16.89
2	1.89	12.4	8.92	24.1
3	0.1	9.84	3.6	23.3
4	0.854	8.7	0.98	27.7
5	8.9	10.3	2.19	36.5

Table 1. SSE for training, validation and test datasets for different M. The specified λ minimizes SSE for validation set given that M.

$SSE_v(M, \lambda)$. In essence we are using the training set to choose the weights, and then using the validation set to optimize over the model, given by (M, λ). For each M we search over $\lambda \in [0, 10]$ to minimize $SSE_v(M, \lambda)$. Then we can check how well our model is doing on the test dataset. Table 3 shows the results.

The result of the model selection is $M = 4, \lambda = 0.85$. However this is misleading. If we plot the three datasets and the prediction function resulting from the weights we get Visually the data look linear. However the one outlier in the training set is leading to bad fits for the $M = 1$ case as can be seen in Figure 8

Hence when we perform our model selection procedure we end up choose $M = 4$ rather than $M = 1$ because this choice of M allows us to roughly hit the outlier and the bulk of the validation data points. However since the test set extends to larger magnitude x values, we have a terrible fit on the test dataset, even though we get a relatively good fit on the validation data. Thus the lesson is that in the presence of outliers model selection can lead us to incorrect results. For the blog feedback there is only one parameter to choose during model selection, namely λ since the feature set Φ has already been specified for us. The objective function that is being minimized during ridge regression is

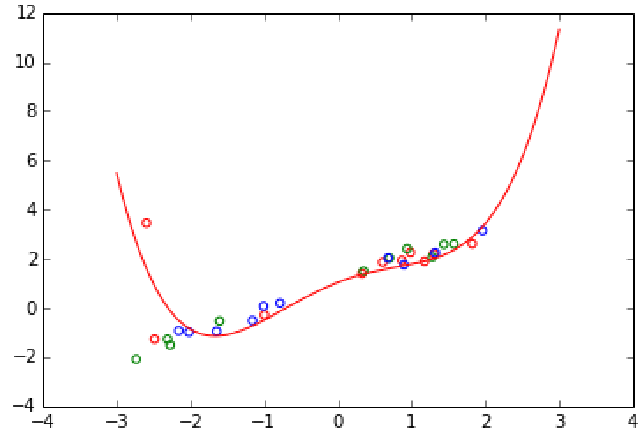

Figure 7. Ridge regression with $M = 3, \lambda = 0.85$. Train datapoints are red, validation are blue, and test are green

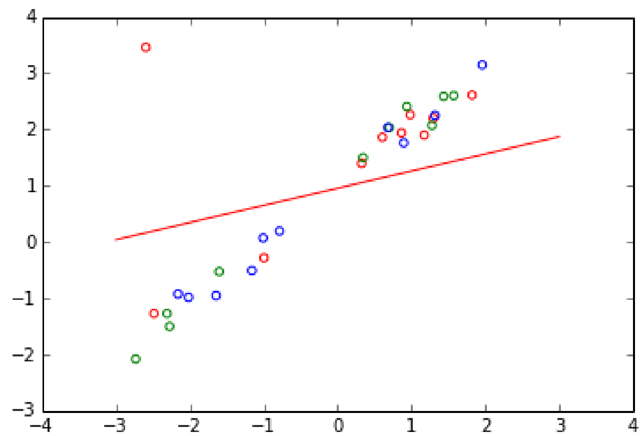

Figure 8. Ridge regression with $M = 1, \lambda = 6.5$. Train datapoints are red, validation are blue, and test are green

$$
\begin{aligned}
&(\Phi w - y)^T(\Phi w - y) + \lambda ||w||_2^2 \\
&= \frac{1}{N}(\Phi w - y)^T(\Phi w - y) + \frac{\lambda}{N}||w||_2^2
\end{aligned}
$$

Hence, it really make sense to think about $\hat{\lambda} = \frac{\lambda}{N}$ since this removes the dependence on the size of the training data. For the curvefitting dataset we had a $\hat{\lambda}$ value of around 0.1. Thus we test $\lambda \in [0.01, 2]$. The results are shown in Figure 9. Interestingly the plot is very flat. So the choice of regularize λ is not having too much of an effect on the MSE of the validation data. In particular it seems that the fit of the model is much less dependent on λ than in the curvefitting example. The plot of $\hat{\lambda}$ against the MSE of the test set looks quite different however. It decreases fairly continuously as $\hat{\lambda}$ increases, see Figure 10. One possibility for this strange behavior could be that the data is not well modeled by the

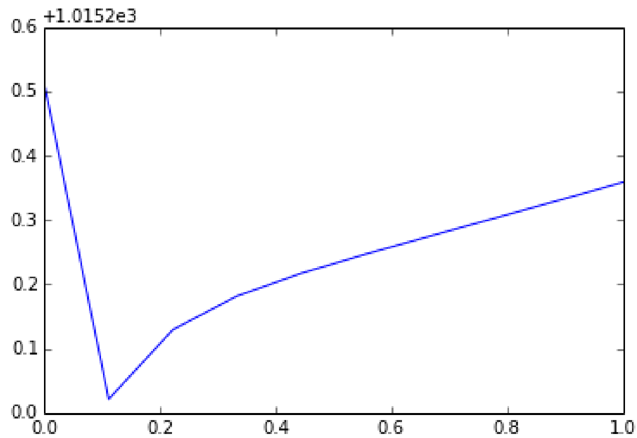

Figure 9. The x-axis is $\hat{\lambda} = \frac{\lambda}{N}$ plotted against the MSE of the validation set on the y-axis.

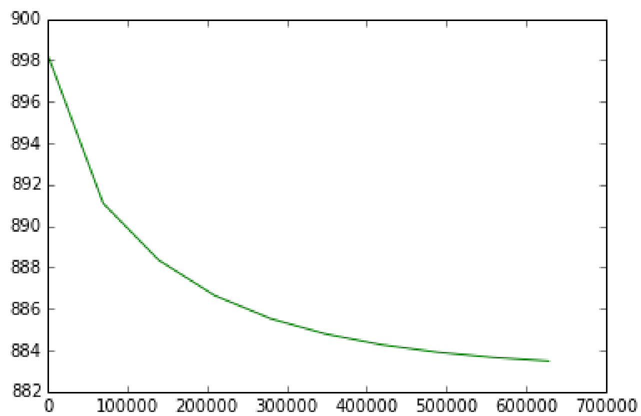

Figure 10. The x-axis is $\hat{\lambda} = \frac{\lambda}{N}$ plotted against the MSE of the test set on the y-axis.

linear model $\hat{y} = w^T x$. Glancing at the data one can see that many of the observations are zero, and then a few have quite a high value, e.g. above 50. If, for example, the true data generating process depends on interactions between features we wouldn't be able to model it well with our linear model. This could be one explanation for the strange performance of the model selection procdedure.

4. Electronic Submission

Submission to ICML 2015 will be entirely electronic, via a web site (not email). Information about the submission process and LaTeX templates are available on the conference web site at:

`http://icml.cc/2015/`

Send questions about submission and electronic templates to `francis.bach@inria.fr`, `david.blei@columbia.edu`.

The guidelines below will be enforced for initial submissions and camera-ready copies. Here is a brief summary:

- Submissions must be in PDF.
- The maximum paper length is **8 pages excluding references, and 10 pages including references** (pages 9 and 10 must contain only references).
- Do **not include author information or acknowledgments** in your initial submission.
- Your paper should be in **10 point Times font**.
- Make sure your PDF file only uses Type-1 fonts.
- Place figure captions *under* the figure (and omit titles from inside the graphic file itself). Place table captions *over* the table.
- References must include page numbers whenever possible and be as complete as possible. Place multiple citations in chronological order.
- Do not alter the style template; in particular, do not compress the paper format by reducing the vertical spaces.

4.1. Submitting Papers

Paper Deadline: The deadline for paper submission to ICML 2015 is at **23:59 Universal Time (3:59 Pacific Daylight Time) on February 6, 2015**. If your full submission does not reach us by this time, it will not be considered for publication. There is no separate abstract submission.

Anonymous Submission: To facilitate blind review, no identifying author information should appear on the title page or in the paper itself. Section 5.3 will explain the details of how to format this.

Simultaneous Submission: ICML will not accept any paper which, at the time of submission, is under review for another conference or has already been published. This policy also applies to papers that overlap substantially in technical content with conference papers under review or previously published. ICML submissions must not be submitted to other conferences during ICML's review period. Authors may submit to ICML substantially different versions of journal papers that are currently under review by the journal, but not yet accepted at the time of submission. Informal publications, such as technical reports or papers in workshop proceedings which do not appear in print, do not fall under these restrictions.

To ensure our ability to print submissions, authors must provide their manuscripts in **PDF** format. Furthermore, please make sure that files contain only Type-1 fonts (e.g., using the program `pdffonts` in linux or using File/DocumentProperties/Fonts in Acrobat). Other fonts (like Type-3) might come from graphics files imported into the document.

Authors using **Word** must convert their document to PDF. Most of the latest versions of Word have the facility to do this automatically. Submissions will not be accepted in Word format or any format other than PDF. Really. We're not joking. Don't send Word.

Those who use **LaTeX** to format their accepted papers need to pay close attention to the typefaces used. Specifically, when producing the PDF by first converting the dvi output of LaTeX to Postscript the default behavior is to use non-scalable Type-3 PostScript bitmap fonts to represent the standard LaTeX fonts. The resulting document is difficult to read in electronic form; the type appears fuzzy. To avoid this problem, dvips must be instructed to use an alternative font map. This can be achieved with the following two commands:

```
dvips -Ppdf -tletter -G0 -o paper.ps paper.dvi
ps2pdf paper.ps
```

Note that it is a zero following the "-G". This tells dvips to use the config.pdf file (and this file refers to a better font mapping).

A better alternative is to use the **pdflatex** program instead of straight LaTeX. This program avoids the Type-3 font problem, however you must ensure that all of the fonts are embedded (use `pdffonts`). If they are not, you need to configure pdflatex to use a font map file that specifies that the fonts be embedded. Also you should ensure that images are not downsampled or otherwise compressed in a lossy way.

Note that the 2015 style files use the `hyperref` package to make clickable links in documents. If this causes problems for you, add `nohyperref` as one of the options to the `icml2015` usepackage statement.

4.2. Reacting to Reviews

We will continue the ICML tradition in which the authors are given the option of providing a short reaction to the initial reviews. These reactions will be taken into account in the discussion among the reviewers and area chairs.

4.3. Submitting Final Camera-Ready Copy

The final versions of papers accepted for publication should follow the same format and naming convention as initial submissions, except of course that the normal author information (names and affiliations) should be given. See Section 5.3.2 for details of how to format this.

The footnote, "Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute." must be modified to "*Proceedings of the 31st International Conference on Machine Learning*, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s)."

For those using the **LaTeX** style file, simply change `\usepackage{icml2015}` to

```
\usepackage[accepted]{icml2015}
```

Authors using **Word** must edit the footnote on the first page of the document themselves.

Camera-ready copies should have the title of the paper as running head on each page except the first one. The running title consists of a single line centered above a horizontal rule which is 1 point thick. The running head should be centered, bold and in 9 point type. The rule should be 10 points above the main text. For those using the **LaTeX** style file, the original title is automatically set as running head using the `fancyhdr` package which is included in the ICML 2015 style file package. In case that the original title exceeds the size restrictions, a shorter form can be supplied by using

`\icmltitlerunning{...}`

just before `\begin{document}`. Authors using **Word** must edit the header of the document themselves.

5. Format of the Paper

All submissions must follow the same format to ensure the printer can reproduce them without problems and to let readers more easily find the information that they desire.

5.1. Length and Dimensions

Papers must not exceed eight (8) pages, including all figures, tables, and appendices, but excluding references. When references are included, the paper must not exceed ten (10) pages. Any submission that exceeds this page limit or that diverges significantly from the format specified herein will be rejected without review.

The text of the paper should be formatted in two columns, with an overall width of 6.75 inches, height of 9.0 inches, and 0.25 inches between the columns. The left margin should be 0.75 inches and the top margin 1.0 inch (2.54 cm). The right and bottom margins will depend on whether you print on US letter or A4 paper, but all final versions must be produced for US letter size.

The paper body should be set in 10 point type with a vertical spacing of 11 points. Please use Times typeface throughout the text.

5.2. Title

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

5.3. Author Information for Submission

To facilitate blind review, author information must not appear. If you are using LaTeX and the `icml2015.sty` file, you may use `\icmlauthor{...}` to specify authors. The author information will simply not be printed until `accepted` is an argument to the style file. Submissions that include the author information will not be reviewed.

5.3.1. Self-Citations

If your are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (**?**), we have shown ...").

Do not anonymize citations in the reference section by removing or blacking out author names. The only exception are manuscripts that are not yet published (e.g. under submission). If you choose to refer to such unpublished manuscripts (**?**), anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look a the Supplementary Material when writing their review.

5.3.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, electronic mail addresses in 10 point small capitals, and physical addresses in ordinary 10 point type. Each author's name should be flush left, whereas the email address should be flush right on the same line. The author's physical address should appear flush left on the ensuing line, on a single line if possible. If successive authors have the same affiliation, then give their physical address only once.

A sample file (in PDF) with author names is included in the ICML2015 style file package.

5.4. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and no more than six or seven sentences.

5.5. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

5.5.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

5.5.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

[1]For the sake of readability, footnotes should be complete sentences.

[2]Multiple footnotes can appear in each column, in the same order as they appear in the text, but spread them across columns and pages if possible.

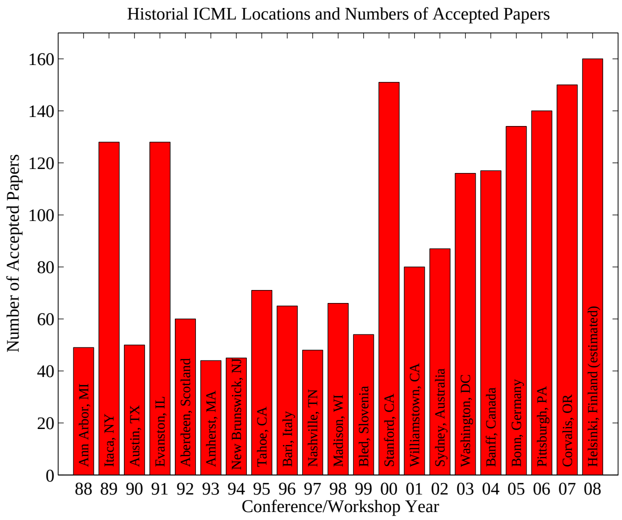

Figure 11. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

5.6. Figures

You may want to include figures in the paper to help readers visualize your approach and your results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 11. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX), but always place two-column figures at the top or bottom of the page.

5.7. Algorithms

If you are using LaTeX, please use the “algorithm” and “algorithmic” environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

Table 2. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9$\pm$ 0.2	96.7$\pm$ 0.2	$\surd$
CLEVELAND	83.3$\pm$ 0.6	80.0$\pm$ 0.6	$\times$
GLASS2	61.9$\pm$ 1.4	83.8$\pm$ 0.7	$\surd$
CREDIT	74.8$\pm$ 0.5	78.3$\pm$ 0.6	
HORSE	73.3$\pm$ 0.9	69.7$\pm$ 1.0	$\times$
META	67.1$\pm$ 0.6	76.5$\pm$ 0.5	$\surd$
PIMA	75.1$\pm$ 0.6	73.9$\pm$ 0.5	
VEHICLE	44.9$\pm$ 0.6	61.5$\pm$ 0.4	$\surd$

5.8. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 2. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material that can be typeset, as contrasted with figures, which contain graphical material that must be drawn. Specify the contents of each row and column in the table’s topmost row. Again, you may float tables to a column’s top or bottom, and set wide tables across both columns, but place two-column tables at the top or bottom of the page.

5.9. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2015.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (**?**). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (**?**). List multiple references separated by semicolons (**???**). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (**?**).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section 5.3 for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (**?**), conference publications (**?**), book chapters (**?**), books (**?**), edited volumes (**?**), technical reports (**?**), and dissertations (**?**).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

5.10. Software and Data

We strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, do not include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look a this material when writing their review.

Acknowledgments

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.