# Messing with linear classifiers

**Gabriel C-Parent**
Département d'informatique et recherche opérationnelle
Université de Montréal
gabriel.c-parent@umontreal.ca

**Dora Fugère**
Département de mathématiques et de statistique
Université de Montréal
dora.fugere@umontreal.ca

## Abstract

Adversarial examples generation from input space in neural network has shown that these powerful constructs can be manipulated into misclassifying previously well classified examples by adding an imperceptible amount of distortion. We wanted to investigate this phenomenon on simpler classifiers.

## 1 Introduction

Recently, neural networks have been brought under questionning. The smoothness assumption, the idea that imperceptible distortion of input shouldn't change the output was shown not to hold [7]. This is a remarkable finding since this is assumed to be a necessary property to the learning process. This comes in stark constrast with feats such as automatic image description [9] and large-scale multi-character text recognition in [3] to name but a few.

As for most real-world problems, there are many desirable and often conflicting goals. Amongst them speed, accuracy and simplicity are easy to justify. We'll focus on comprehensibility, because that justifies our focus on a simpler model. We interpet simplicity as "given two models with the same generalization error, the more comprehensible one should be preferred" [2]. This obviously is dependent on multiple other factors (e.g. speed and accuracy) but it does sound like the *keep it simple stupid* rule of thumb. Furthermore, as stated in [4], empirical comparison of performance is very context-dependent and can be infludenced by treatments such as the preprocessing steps, training parameters and model hyperparameters.

Inspired by the methodology to induce misclassification, we wondered if a similar optimization procedure could be applied to generate adversarial examples in a simpler classifier. For this purpose, we used a linear support vector machine (SVM). The only other contender would have been Naive Bayes, but we happen to like sklearn's implementation of the linear SVM [1] [6]. We report the robustness of our optimization procedure, the results on classifiers trained with different loss and regularization parameters and we then feed the generated adversarial examples to a neural network to see if some underlying feature of the image was captured.

[1] we wouldn't risk reinventing the square wheel

## 2 Framework

### 2.1 Dataset

The experiments were performed on the MNIST dataset [5].

Let $X = \{0, 255\}^{784}$, the input domain. This is the set of $28 \times 28$ 8-bit image.

Let $Y = \{0, 9\}$, the output domain. This is the set of valid classes for an MNIST digit.

### 2.2 Optimization goal

Let $f : X \to Y$ a classifier mapping $x_i \in X$ to $y_i \in Y$.

We aim to solve the following optimization

$$\begin{aligned} \text{minimize} \quad & \|r\|^2 \\ \text{subject to} \quad & \\ & x_i + r \in X \\ & f(x_i) \neq f(x_i + r) \end{aligned} \tag{1}$$

This is quite similar to [7] but the newly generated images remain 8-bit to stay in the input domain of the MNIST dataset. Sadly, this also makes it a discrete optimization problem.

### 2.3 Optimization goal for the linear SVM

Suppose we want to misclassy an arbitrary image $x_i$ correctly classified as $y_1$ by adding a vector $r$ of distortion in a two-class setting.

The classifier classifies the input based on the following decision function.

$$y_i = argmax(x_i \cdot W^T + b) \tag{2}$$

The difference between the class weights of the classifier is $W_d iff$.

$$W_{diff} = W_2 - W_1 \tag{3}$$

The distance between the values of the two classes is $d$.

$$d = x_i \cdot W_{diff}^T \tag{4}$$

To cause misclassification, $r$ must respect the following constraint:

$$r \cdot W_{diff}^T > d \tag{5}$$

What we need is to find the smallest $\|r\|^2$ that will cause misclassification. Note that when there are more than two classes, we just apply the procedure to all other classes $y_i \neq y_1$ and choose the one with minimal squared euclidean norm.

### 2.4 Knapsack problem and the greedy approach

The problem is similar to the bounded multiple-class binary Knapsack problem [8], with the difference that we are searching for the smallest knapsack holding a value superior to $d$ (equation 4).

The exact algorithm would be too costly for our purpose so we chose to use a greedy heuristic inspired by Dantzig's [1].

We encode the cost of adding a bit of distortion to a pixel in the form of a volume. For example, given vector of distortion $r$ = [0, 3], the costs of adding 1 unit of distortion for each position of $r$ is

[1, 7]. For the values, we use the absolute value of vector $W_{diff}$. If $W_{diff}$ = [-5, 10], the $\frac{value}{volume}$ ratios would be [5, $\frac{10}{7}$]. The best choice to improve the value

We iteratively choose the elements with the best until the critical value is $d$ is passed.

At iteration $n$, the choices made are optimal of the knapsack of size $K_n = \sum_{i=1}^{n} w_i$. This is the case because every object in the knapsack was chosen with maximum $\frac{value}{volume}$ ratio and the knapsack is filled perfectly.

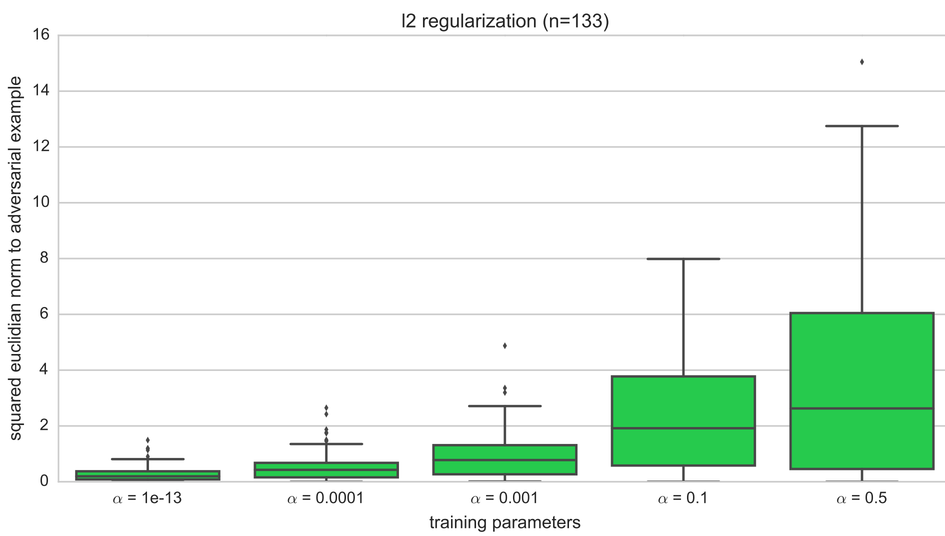
Let $(K_1, v_1)$ and $(K_2, v_2)$, the size and value of knapsacks at two iterations, we know that if $K_1 < K_2$ then $v_1 \leq v_2$, that is to say it's a monotonically increasing function. The inverse is also true, if $v_1 < v_2$ then $K_1 < K_2$.

For our purpose, this means that the true value of the distance $\|r\|^2 \in [K_{n-1}, K_n]$ if the algorithm finishes at iteration $n$. This boundary can be very small or very big, depending on the size of the weight increments. We measured the boundary size to show that in this problem, we have very little incertitude.

# 3 Experimental results

## 3.1 Interval of $\|r\|^2$

## 3.2 Regularization schemes




## 3.3 Style

Papers to be submitted to NIPS 2014 must be prepared according to the instructions presented here. Papers may be only up to eight pages long, including figures. Since 2009 an additional ninth page *containing only cited references* is allowed. Papers that exceed nine pages will not be reviewed, or in any other way considered for presentation at the conference.

Please note that this year we have introduced automatic line number generation into the style file (for LaTeX $2_\varepsilon$ and Word versions). This is to help reviewers refer to specific lines of the paper when they make their comments. Please do NOT refer to these line numbers in your paper as they will be removed from the style file for the final version of accepted papers.

The margins in 2014 are the same as since 2007, which allow for $\approx 15\%$ more words in the paper compared to earlier years. We are also again using double-blind reviewing. Both of these require the use of new style files.

Authors are required to use the NIPS LaTeX style files obtainable at the NIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

### 3.4 Retrieval of style files

The style files for NIPS and other conference information are available on the World Wide Web at

`http://www.nips.cc/`

The file `nips2014.pdf` contains these instructions and illustrates the various formatting requirements your NIPS paper must satisfy. LaTeX users can choose between two style files: `nips11submit_09.sty` (to be used with LaTeX version 2.09) and `nips11submit_e.sty` (to be used with LaTeX2e). The file `nips2014.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own. The file `nips2014.rtf` is provided as a shell for MS Word users.

The formatting instructions contained in these style files are summarized in sections 4, 5, and 6 below.

## 4 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing of 11 points. Times New Roman is the preferred typeface throughout. Paragraphs are separated by 1/2 line space, with no indentation.

Paper title is 17 point, initial caps/lower case, bold, centered between 2 horizontal rules. Top rule is 4 points thick and bottom rule is 1 point thick. Allow 1/4 inch space above and below title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in section 6 regarding figures, tables, acknowledgments, and references.

## 5 Headings: first level

First level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 12. One line space before the first level heading and 1/2 line space after the first level heading.

### 5.1 Headings: second level

Second level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the second level heading and 1/2 line space after the second level heading.

#### 5.1.1 Headings: third level

Third level headings are lower case (except for first word and proper nouns), flush left, bold and in point size 10. One line space before the third level heading and 1/2 line space after the third level heading.

## 6 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

### 6.1 Citations within the text

Citations within the text should be numbered consecutively. The corresponding number is to appear enclosed in square brackets, such as [1] or [2]-[5]. The corresponding references are to be listed in the same order at the end of the paper, in the **References** section. (Note: the standard BIBTEX style `unsrt` produces this.) As to the format of the references themselves, any style is acceptable as long as it is used consistently.

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]", not "In our previous work [4]". If you cite your other papers that are not widely available (e.g. a journal paper under review), use anonymous author names in the citation, e.g. an author of the form "A. Anonymous".

### 6.2 Footnotes

Indicate footnotes with a number[2] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[3]

### 6.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

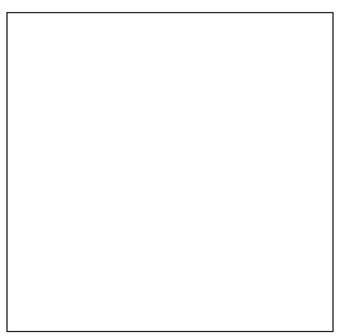


Figure 1: Sample figure caption.

### 6.4 Tables

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 1.

[2] Sample of the first footnote
[3] Sample of the second footnote

Table 1: Sample table title

| PART | DESCRIPTION |
|---|---|
| Dendrite | Input terminal |
| Axon | Output terminal |
| Soma | Cell body (contains cell nucleus) |

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

## 7 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

## 8 Preparing PostScript or PDF files

Please prepare PostScript or PDF files with paper size "US Letter", and not, for example, "A4". The -t letter option on dvips will produce US Letter files.

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- LaTeX users:
  - Consider directly generating PDF files using `pdflatex` (especially if you are a MiKTeX user). PDF figures must be substituted for EPS figures, however.
  - Otherwise, please generate your PostScript and PDF files with the following commands:

    ```
    dvips mypaper.dvi -t letter -Ppdf -G0 -o mypaper.ps
    ps2pdf mypaper.ps mypaper.pdf
    ```

    Check that the PDF files only contains Type 1 fonts.
  - xfig "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
  - The `\bbold` package almost always uses bitmap fonts. You can try the equivalent AMS Fonts with command

    ```
    \usepackage[psamsfonts]{amssymb}
    ```

    or use the following workaround for reals, natural and complex:

    ```
    \newcommand{\RR}{I\!\!R} %real numbers
    \newcommand{\Nat}{I\!\!N} %natural numbers
    \newcommand{\CC}{I\!\!\!\!C} %complex numbers
    ```
  - Sometimes the problematic fonts are used in figures included in LaTeX files. The ghostscript program `eps2eps` is the simplest way to clean such figures. For black and white figures, slightly better results can be achieved with program `potrace`.

- MSWord and Windows users (via PDF file):
  - Install the Microsoft Save as PDF Office 2007 Add-in from `http://www.microsoft.com/downloads/details.aspx?displaylang=en&familyid=4d951911-3e7e-4ae6-b059-a2e79ed87041`
  - Select "Save or Publish to PDF" from the Office or File menu
- MSWord and Mac OS X users (via PDF file):
  - From the print menu, click the PDF drop-down box, and select "Save as PDF..."
- MSWord and Windows users (via PS file):
  - To create a new printer on your computer, install the AdobePS printer driver and the Adobe Distiller PPD file from `http://www.adobe.com/support/downloads/detail.jsp?ftpID=204` *Note:* You must reboot your PC after installing the AdobePS driver for it to take effect.
  - To produce the ps file, select "Print" from the MS app, choose the installed AdobePS printer, click on "Properties", click on "Advanced."
  - Set "TrueType Font" to be "Download as Softfont"
  - Open the "PostScript Options" folder
  - Select "PostScript Output Option" to be "Optimize for Portability"
  - Select "TrueType Font Download Option" to be "Outline"
  - Select "Send PostScript Error Handler" to be "No"
  - Click "OK" three times, print your file.
  - Now, use Adobe Acrobat Distiller or ps2pdf to create a PDF file from the PS file. In Acrobat, check the option "Embed all fonts" if applicable.

If your file contains Type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

### Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

### References

## References

[1] DANTZIG, G. B. Discrete-variable extremum problems. *Operations Research 5*, 2 (1957).

[2] DOMINGOS, P. The role of occam's razor in knowledge discovery. *Data mining and knowledge discovery 3*, 4 (1999), 409–425.

[3] GOODFELLOW, I. J., BULATOV, Y., IBARZ, J., ARNOUD, S., AND VINAY, S. Multi-digit number recognition from street view.

[4] HAND, D. J. Classifier technology and the illusion of progress. *Statistical Science 21*, 1 (Feb. 2006), 1–14.

[5] LECUN, Y., AND CORTES, C. The mnist database of handwritten digits.

[6] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[7] SZEGEDY, C., ZAREMBA, W., SUTSKEVER, I., BRUNA, J., ERHAN, D., GOODFELLOW, I., AND FERGUS, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[8] VANDERBECK, F. Extending dantzig's bound to the bounded multiple-class binary knapsack problem. *Mathematical Programming 94*, 1 (Dec. 2002), 125–136.

[9] VINYALS, O., TOSHEV, A., BENGIO, S., AND ERHAN, D. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555* (2014).