# Draft

22. maj 2014

## 1 Abstract

Ma et al. [1] has shown leverage sampling to outperform uniform sampling for Least-Squares regression. We explore the possibility of using the same sampling distribution on 2-class classification, and introduce a new leverage distribution based on a generalization of the idea.

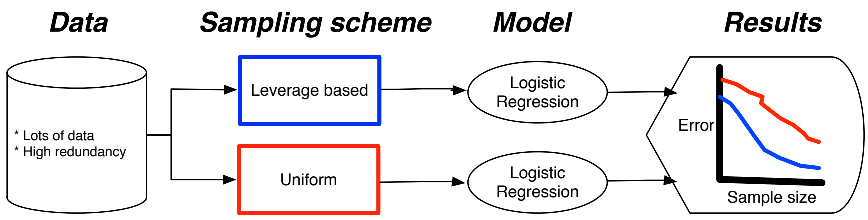
## 2 Motivation

For video the importance of sampling methods is exemplified by very large and high-dimensional datasets where

- It is not feasible to use all of the available data at once.
- There is a high redundancy between datapoints (25 fps).
- Computational cost is rarely linear to the input size.

We therefore want to explore alternative sampling methods, and try to identify datapoints which are important when fitting a model.

## 3 Concept



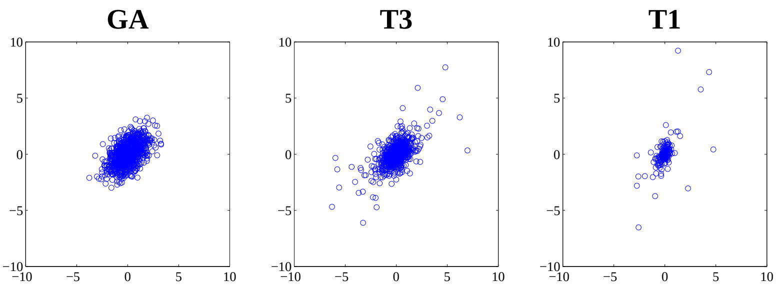
## 4 Research Questions

- Can we validate the results for least-squares regression shown by Ma et al. ?
- Will a linear regression based sampling distribution improve our performance in classification?
- Can leverage based sampling be generalized and used for classification?

## 5 Datasets

These datasets are drawn from distributions defined in Ma et al. [**?**] and characterised by

- GA: Nearly uniform leverage-scores
- T3: Mildly non-uniform leverage-scores
- T1: Very non-uniform leverage-scores




**Figur 1:** The three distributions considered standardized for comparison

## 6 Leveraging for least-squares regression

When fitting a model, we know that some datapoints are more important that others, leveraging is based on the idea that we can determine the importance of these point beforehand.

1. A leverage-score is calculated for each datapoint (its importance).
2. These scores are normalized into a distribution $\pi$ to sample from.

Ma. et al. [?] use the leverage-scores for least-square regression defined as the diagonal elements of

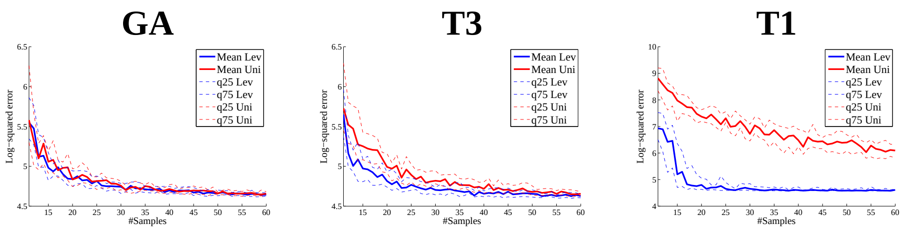
$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{6.1}$$

This comes from the closed form expression for predictions which is linear in $y$

$$\hat{\mathbf{y}}_n = \mathbf{X}_n * \hat{\beta} \quad \text{where} \quad \hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

# 7 Validation of the results Ma et al.

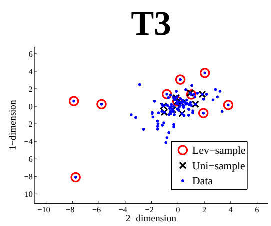
We have empirically tested and validated the results shown by Ma et al. [?].




**Figur 2:** Comparison of uniform **(red)** vs. leverage **(blue)** based sampling schemes for least-squares regression. $N = 1000$, $d = 10$.

- GA: The leverage score are approximately uniform, and thus there is no significant difference between the two sampling schemes.
- T3: Leveraging consistently provides slightly better results compared to uniform sampling.
- T1: With *very non-uniform* leverage-scores, leveraging clearly outperforms uniform sampling.

**Figur 3:** Comparison of sampling methods



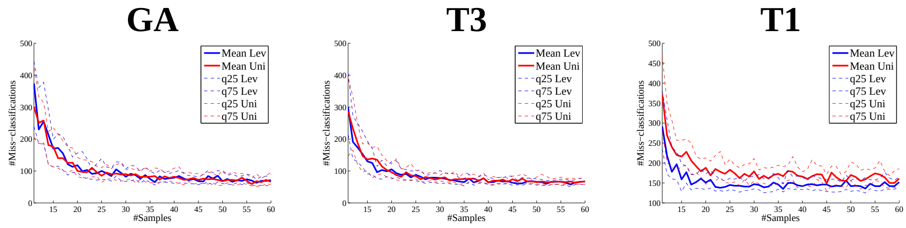

There results are consistent when varying $N$ and $d$, although the level of improvement varies.

# 8 LS-based Distribution for Classification

We sample from the same distribution (6.1) as for least-squares regression. We use these samples to train a logistic regression model for 2 class classification, with equal class size.

# 9 Test Results

We compared the LS-distribution **(blue)** to a uniform-distribution **(red)** in sampling for a logistic regression. The mean, 25th and 75th quantile are plotted.




- Sampling from the LS-distribution is no better that uniform on datasets of type GA and T3.
- With very non-uniform leverage scores, T1, the LS-distribution slightly outperforms uniform sampling.

The results shown are for dimension $p = 10$ and $N = 1000$ datapoints, but it is consistent when varying $p$ and $N$.

# 10 Sensitivity Based Distribution

We generalize the leverage scores to other models by seeing that they can be described as:

$$\frac{\delta \hat{\mathbf{y}}_n}{\delta \mathbf{y}_n} = Diag\,(H) \tag{10.1}$$

Which we call the sensitivity of the model to a specific datapoint. For a general probabilistic discriminative model this requires the following:

$$\hat{\mathbf{y}}_n = p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}}) \quad \bar{\mathbf{w}} \text{ s.t. } \frac{\delta L}{\delta \bar{\mathbf{w}}} = 0 \tag{10.2}$$

Since 15.3 depends both directly and indirectly on $y$ we see that

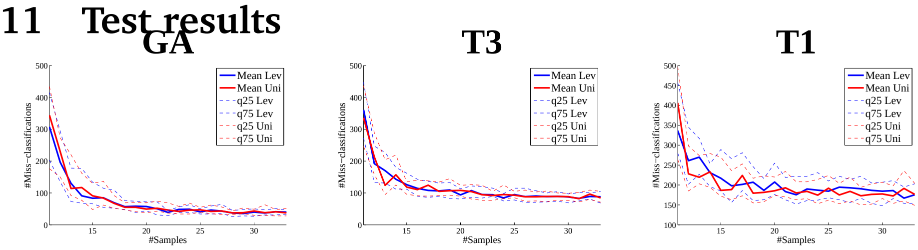
$$\frac{\delta}{\delta \mathbf{y}} \frac{\delta \mathcal{L}}{\delta \mathbf{w}} = 0 \Rightarrow \frac{\delta^2 \mathcal{L}}{\delta \mathbf{y} \delta \bar{\mathbf{w}}} + \frac{\delta^2 \mathcal{L}}{\delta \bar{\mathbf{w}} \delta \bar{\mathbf{w}}^T} \frac{\delta \bar{\mathbf{w}}}{\delta \mathbf{y}} = 0 \tag{10.3}$$

and from this we can get our leverage-score (15.1)

$$\frac{\delta \hat{\mathbf{y}}_n}{\delta \mathbf{y}_n} = \frac{\delta p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}})}{\delta \bar{\mathbf{w}}^T} \frac{\delta \bar{\mathbf{w}}}{\delta \mathbf{y}} = -\frac{\delta p(y|\bar{\mathbf{x}}_n, \bar{\mathbf{w}})}{\delta \bar{\mathbf{w}}^T} \left[ \frac{\delta^2 \mathcal{L}}{\delta \bar{\mathbf{w}} \delta \bar{\mathbf{w}}^T} \right]^{-1} \frac{\delta^2 \mathcal{L}}{\delta \mathbf{y} \delta \bar{\mathbf{w}}}$$

When using this model, initial weights are found by fitting a small uniform sample. This is expected outperform LS-based sampling since it introduces dependence on class information.

## 11 Test results




**Figur 4:** Comparison of sensitivity vs. uniform -based sampling for logistic regression.

We see that the *sensitivity based sampling* gives us a performance equivalently to that of uniform sampling.

## 12 Future work

From our work several new question arise.

- How large show the initial sampling size be for sensitivity-based sampling?
- How should the non-linear sensitivity based leverage scores be normalised?
- Should all points be sampled from the initial weights found, or should the process be iterative?

## 13 Conclusion

In the case of linear regression, leverage-based sampling provides a improvement over uniform sampling when the leverage-scores are mildly or very non-uniform.

Using the LS-based sampling for classification is slightly better with very non-uniform leverage-scores, T1 data.

We have generalized the concept of leverage-based scores to classification with logistic regression and it has shown no improvements. However further analysis and tweaking might improved this approach.

## 14 References

## 15 The derivative of prediction or Sensitivity

We wish to find the effect that a datapoint's class has on the predicted class for that datapoint.

$$\frac{\delta \hat{Y}_n}{\delta Y_n} \tag{15.1}$$

Our prediction is

$$\hat{Y}_n = p(y|\bar{x}, \bar{w}) \tag{15.2}$$

where $\bar{w}$ is subject to

$$\frac{\delta L}{\delta \bar{w}} = 0 \tag{15.3}$$

Which means that we have found a locally optimal solution.

We now assume that when we move $y$ by a small amount $\delta y$ then 15.3 still holds. (can we do this with a discrete y ?)

Essentially assuming some smoothness around the optimum.

Using this and the fact that 15.3 depends both directly and indirectly on y we see that

$$\frac{\delta}{\delta y}\frac{\delta L}{\delta w} = 0$$
$$\Downarrow$$
$$\frac{\delta^2 L}{\delta y \delta \bar{w}} + \frac{\delta^2 L}{\delta \bar{w} \delta \bar{w}^T}\frac{\delta \bar{w}}{\delta y} = 0$$

and from this we can isolate

$$\frac{\delta \bar{w}}{\delta y} = -\left[\frac{\delta^2 L}{\delta \bar{w}\delta \bar{w}^T}\right]^{-1}\frac{\delta^2 L}{\delta y \delta \bar{w}} \tag{15.4}$$

Rewriting (15.1) we get

$$\frac{\delta \hat{Y}_n}{\delta Y_n} = \frac{\delta p(y|\bar{x}_n, \bar{w})}{\delta Y_n} = \frac{\delta p(y|\bar{x}_n, \bar{w})}{\delta \bar{w}^T}\frac{\delta \bar{w}}{\delta y} \tag{15.5}$$

And inserting (15.4)

$$\frac{\delta p(y|\bar{x}_n, \bar{w})}{\delta \bar{w}^T}\frac{\delta \bar{w}}{\delta y} = -\frac{\delta p(y|\bar{x}_n, \bar{w})}{\delta \bar{w}^T}\left[\frac{\delta^2 L}{\delta \bar{w}\delta \bar{w}^T}\right]^{-1}\frac{\delta^2 L}{\delta y \delta \bar{w}} \tag{15.6}$$

And this is our leverage score for this where $\bar{w}$ is gained from a training based on a small dataset.

# 16 Randomised algorithm

*Uncertainty based on asymptotic likelihood and $\bar{w}$-distribution*

Let $\mathcal{L}_\infty$ be the log-likelihood function for a distribution, now let $\mathcal{L}_N$ denote the log-likelihood function based on $N$ observations from this distribution. Furthermore, let $N$ be a large number, for which $L_N \approx L_\infty$.

$$\mathcal{L}_{\mathcal{N}} = \frac{1}{N}\sum_{n=1}^{N}\ell_n \qquad \bar{w}\; s.t. \frac{\delta \mathcal{L}}{\delta \bar{w}} = \bar{0} \tag{16.1}$$

Where $\ell_n$ is the log-likelihood of the $n^{th}$ observation. And $\bar{w}$ is the true weights for the distribution, then we combine the expressions from (16.1), such that for the true weights the following must be fulfilled:

$$\frac{1}{N}\sum_{n=1}^{N}\frac{\delta \ell_n}{\delta \bar{w}} = 0 \tag{16.2}$$

*(Skal vi lige skrive lidt om at $\Delta w = w - w_0$ og er en lille forskydelse i vægtene? Eller er det en lille forskydelse?)* For each of the $N$ observations, we can approximate the log-likelihood of the $n^{th}$ observation with this taylor expansion:

$$\ell_n(\bar{w}) = \ell_n(\bar{w}_0) + \left.\frac{\delta \ell_n}{\delta \bar{w}}\right|_{\bar{w}_0} \Delta\bar{w} + \frac{1}{2} Tr\left[\left.\frac{\delta \ell_n}{\delta \bar{w}\delta \bar{w}^T}\right|_{\bar{w}_0} \Delta\bar{w}\Delta\bar{w}^T\right] \tag{16.3}$$

Or for the entire log-likelihood function: **Where did the trace go ?**

$$\mathcal{L}_N(\bar{w}) = \mathcal{L}_N(\bar{w}_0) + \left(\left.\frac{\delta \mathcal{L}_N}{\delta \bar{w}}\right|_{\bar{w}_0}\right)^T \cdot \Delta\bar{w} + \frac{1}{2}\Delta\bar{w}^T\left(\left.\frac{\delta^2 \mathcal{L}_N}{\delta \bar{w}\delta \bar{w}^T}\right|_{\bar{w}_0}\right)\Delta\bar{w} + R \tag{16.4}$$

Where $R$ is the error of the approximation and assumed to be 0. Furthermore, we define $\bar{\bar{H}}_N = \left.\frac{\delta^2 \mathcal{L}_N}{\delta \bar{w}\delta \bar{w}^T}\right|_{\bar{w}_0}$, and $\bar{g} : N = \left.\frac{\delta \mathcal{L}_N}{\delta \bar{w}}\right|_{\bar{w}_0}$. And evaluate condition (16.1) on $\bar{w}$:

$$\frac{\delta \mathcal{L}_{\mathcal{N}}}{\delta \bar{w}} = \bar{g}_N + \bar{\bar{H}}_N \Delta\bar{w} = \bar{0} \tag{16.5}$$

We replace $\Delta\bar{w}$ with $\Delta\hat{\bar{w}}$ as $N$ is a finite number, thus only approximating $\Delta\bar{w}$. Isolating $\Delta\hat{\bar{w}}$, and using Ljung **[REFERENCE?]**, we get: **Is this to soon to involve Ljung?**

$$\Delta\hat{\bar{w}} = -\bar{\bar{H}}_N^{-1} \cdot \bar{g}_N \overbrace{=}^{Ljung} -\bar{\bar{H}}_0^{-1} \cdot \bar{g}_{\bar{w}}(\bar{w}_0) \tag{16.6}$$

*(Forklaring af at $H_0$ er uafhængig af datasæt, mens $g$ nu er afhængig af $w$ evalueret i $w_0$)* Besides getting an estimate for $\Delta\hat{\bar{w}}$, we can find the mean of the distribution:

$$\left\langle \Delta\hat{\bar{w}} \right\rangle = -\bar{\bar{H}}_0^{-1}\bar{g}_0 = 0$$

As $\delta\bar{w} = \bar{w} - \bar{w}_0$ ??mistet trraden?

## 16.1 Covariance of $\bar{w}$ - distribution

**Why do we do this???**

$$\left\langle \delta\bar{w}\delta\bar{w}^T \right\rangle_N = \left\langle \bar{\bar{H}}^{-1}\bar{g}\bar{g}^T\bar{\bar{H}}^{-1} \right\rangle \overbrace{=}^{Ljung} \bar{\bar{H}}_0^{-1}\left\langle \bar{g}\bar{g}^T \right\rangle \bar{\bar{H}}_0^{-1} + R' \tag{16.7}$$

With error $R' = O\left(\frac{1}{N}\right) \approx 0$, for large $N$. We look at the covariance of the gradient function

$$\left\langle \bar{g}\bar{g}^T \right\rangle_N = \frac{1}{N^2} \sum_{n,n'=1}^{N} \left\langle \left.\frac{\delta \ell_n}{\delta_n \bar{w}}\right|_{\bar{w}_0} \left.\frac{\delta \ell_{n'}}{\delta_n \bar{w}}\right|_{\bar{w}_0} \right\rangle \tag{16.8}$$

$$= \frac{1}{N^2} \left( \sum_{n \neq n'} \underbrace{\left\langle \left.\frac{\delta \ell_n}{\delta \bar{w}}\right|_{\bar{w}_0} \right\rangle \cdot \left\langle \left.\frac{\delta \ell_{n'}}{\delta \bar{w}^T}\right|_{\bar{w}_0} \right\rangle}_{0} + \sum_{n=1}^{N} \left\langle \left.\frac{\delta \ell_n}{\delta \bar{w}}\right|_{\bar{w}_0} \left.\frac{\delta \ell_n}{\delta \bar{w}^T}\right|_{\bar{w}_0} \right\rangle \right) \tag{16.9}$$

Due to the assumption of independence, only the $N$ diagonal elements are non-zero. So;

$$\left\langle \bar{g}\bar{g}^T \right\rangle_N = \frac{1}{N} \left\langle \left.\frac{\delta \mathcal{L}}{\delta \bar{w}}\right|_{\bar{w}_0} \left.\frac{\delta \mathcal{L}}{\delta \bar{w}^T}\right|_{\bar{w}_0} \right\rangle \tag{16.10}$$

## 16.2 Proof that $\left\langle \left.\frac{\delta \mathcal{L}}{\delta \bar{w}}\right|_{\bar{w}_0} \left.\frac{\delta \mathcal{L}}{\delta \bar{w}^T}\right|_{\bar{w}_0} \right\rangle = \bar{\bar{H}}_0$

Tekst test

$$\left\langle \bar{g}\bar{g}^T \right\rangle_N = \frac{1}{N^2} \sum_{n=1}^{N} \int_\Omega \left.\frac{\delta \ell_n(\bar{x})}{\delta \bar{w}}\right|_{\bar{w}_0} \left.\frac{\delta \ell_n(\bar{x})}{\delta \bar{w}}\right|_{\bar{w}_0} p(\bar{x})\delta x \tag{16.11}$$

From (16.10) and (16.11), and setting $\ell_n(\bar{x}) = p(\bar{x})$:

$$\left.\bar{\bar{H}}\right|_{\bar{w}_0} = \frac{1}{N} \sum_{n=1}^{N} \int_\Omega \frac{\delta}{\delta \bar{w} \delta \bar{w}^T} - \log p\left(\bar{x}|\bar{w}\right) p(\bar{x})\delta x \tag{16.12}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \int_\Omega -\frac{\delta}{\delta \bar{w}} \frac{1}{p\left(\bar{x}\right)} \frac{\delta}{\delta \bar{w}^T} p(\bar{x}|\bar{w}) p(\bar{x}) \delta \bar{x} \tag{16.13}$$

$$\tag{16.14}$$

Now if $p(\bar{x}|\bar{w}_0) = p(x)$, then

## 16.3 Uncertainty of prediction

For a number of weight-vectors $\bar{w}$, we take the mean of predictions based on these weight-vectors;

$$\left\langle p\left(y|\bar{x}, \bar{w}\right) \right\rangle \approx p\left(y \middle| \bar{x}, \hat{\bar{w}}\right) = p\left(y|\bar{x}, \mathbf{E}(\bar{w})\right) \tag{16.15}$$

We now look at a small change in the prediction $\Delta p$, caused by a change of $\Delta \bar{w}$ in true weight vector $\bar{w}_0$.

$$\Delta p = p\left(y|\bar{x}, \bar{w}_0 + \Delta \bar{w}\right) - p\left(y|\bar{x}, \bar{w}_0\right) \approx \left.\frac{\delta p}{\delta \bar{w}}\right|_{w_0} \cdot \Delta \bar{w} \tag{16.16}$$

The variance of $\Delta p$, can then be computed as

$$\left\langle (\Delta p)^2 \right\rangle = Tr\left[\frac{\delta p}{\delta \bar{w}}\left(\frac{\delta p}{\delta \bar{w}}\right)^T \left\langle \Delta \bar{w} \Delta \bar{W}^T \right\rangle\right] = \frac{1}{N}\left(\frac{\delta p}{\delta \bar{w}}\right)^T \bar{\bar{H}}^{-1}\frac{\delta p}{\delta \bar{w}} \quad (16.17)$$

### 16.3.1 For a linear model with known $\sigma^2$

The prediction in a linear model is:

$$p(y|\bar{x},\bar{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y-f(\bar{x}||\bar{w})}{2\sigma^2}} \quad (16.18)$$

Where $y$ is the target and $f(\bar{x}||\bar{w})$ is the prediction. Differentiating (**??**) with respect to $\bar{w}$: *(Hvorfor er det vi gør det??)*

$$\frac{\delta p}{\delta w} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-f(\bar{x}|\bar{w})^2}{2\sigma^2}} - (y - f(\bar{x}|\bar{w})\frac{\delta f(\bar{x}|\bar{w})}{\delta \bar{w}} \quad (16.19)$$

We let $y = f(\bar{x}|\bar{w}) + \epsilon$. (Targets kan beskrives som en approximativ funktion + en fejl ..)

$$\frac{\delta p}{\delta \bar{w}} = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^2}{2\sigma^2}}\epsilon^2}_{\texttt{const. w.r.t. } \bar{x}} \frac{\delta f(\bar{x}|\bar{w})}{\delta \bar{w}}^T \bar{\bar{H}}_0^{-1}\frac{\delta f(\bar{x}|\bar{w})}{\delta \bar{w}} \quad (16.20)$$