Appendix

Experimental Simulation of Example 1

Here we experimentally simulate Example 1 to illustrate that logistic regression classifier has large l_1 error. We use Latent Dirichlet Allocation (LDA) (Blei et al., 2003), the state of the art generative model for documents, to generate datasets. The detailed experiment settings are listed below:

- The dataset consists of 20000 documents, the number of topics is 20, the dictionary size is 1000, and the average number of words in each document is 200.
- We use the non-informative Dirichlet prior $\alpha = (1, 1, \ldots, 1)$ over topics. The word distribution in each topic follows power law with a random order among words.
- For each document, we randomly sample with replacement 10 topic labels from the topic distribution.

Table 1 reports the mean experiment results and the standard deviation across five runs. For reference we also include the relative frequency of labels, and the l_1 error achieved by the trivial classifier that always output the global relative frequency of labels as conditional probability.

Average l_1 Error	Empirical Calibration
0.1270 ± 0.0008	0.0083 ± 0.0003
Trivial l_1 Error	Frequency of Labels
0.2022 ± 0.0001	0.3448 ± 0.0001

Table 1: L_1 error and empirical calibration

As we can see from Table 1, the logistic regression only achieves 0.13 average l_1 error, while even the trivial classifier can achieve 0.2. This implies that logistic regression performed very badly in this example. However, as we can see from Table 1, the empirical calibration measure of logistic regression classifier is relatively low (0.01), indicating that the classifier is almost calibrated.

Proof of Theorem 1

Proof. The proof relies on the following lemma:

Lemma 1. *Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. Let D be a size n i.i.d. sample set from $\mathcal{P}$. Let V be a verifier of $\mathcal{P}$ given D (i.e., V is a function from $\{\mathcal{X} \times \mathcal{Y}\}^n$ to $\{0, 1\}$), such that*

1. *With probability at least $1 - \delta_1$, a dataset D with n i.i.d. samples from $\mathcal{P}$ will pass V:*

$$\mathbf{Pr}_D(V(D) = 1) \geq 1 - \delta_1$$

2. *With probability at least $1 - \delta_2$, a dataset D with n i.i.d. samples from $\mathcal{P}$ satisfies:*

$$\mathbf{Pr}(\forall i \neq j, X_i \neq X_j) \geq 1 - \delta_2$$

Then there exists another probability distribution $\mathcal{P}'$ such that:

1. *With probability at least $1-\delta_1-\delta_2$, a data D' with n i.i.d. samples from $\mathcal{P}'$ will also pass V.*

$$\mathbf{Pr}_{D'}(V(D') = 1) \geq 1 - \delta_1 - \delta_2$$

2.
$$\forall X \in \mathcal{X}, \sum_{Y \in \mathcal{Y}} \mathcal{P}(X, Y) = \sum_{Y \in \mathcal{Y}} \mathcal{P}'(X, Y)$$

3.
$$\forall X \in \mathcal{X}, \mathcal{P}'(Y = 1|X) = 0 \text{ or } 1$$

Proof. First we construct the following distribution over all possible $\mathcal{P}'$ satisfying the last two conditions:

$$\mathbf{Pr}(\mathcal{P}') = \prod_{X \in \mathcal{X}} Q(P'(Y = 1|X), P(Y = 1|X))$$

where $Q(p', p)$ is defined as:

$$Q(p', p) = \begin{cases} p & p' = 1 \\ 1 - p & p' = 0 \end{cases}$$

Now it is sufficient to show that if we sample $\mathcal{P}'$ according to the above distribution and then sample D' from $\mathcal{P}'$, then with probability at least $1 - \delta_1 - \delta_2$, D' will pass V. Assuming this is true, then at least one distribution $\mathcal{P}'$ have to satisfy the first condition, and thereby proved the existence of $\mathcal{P}'$.

To compute the probability that D' would pass V, denote $D_X = \{X_1, X_2, \ldots, X_n\}$ and $D_Y = \{Y_1, Y_2, \ldots, Y_n\}$. Note that all $\mathcal{P}'$ has the same marginal distribution over $\mathcal{X}$, therefore:

$$\mathbf{Pr}_{\mathcal{P}', D'}(V(D') = 1) = \sum_{\mathcal{P}'} \mathbf{Pr}(\mathcal{P}') \sum_{D'} \mathbf{Pr}(D'|\mathcal{P}') V(D')$$
$$= \sum_{D'_X} \mathbf{Pr}(D'_X) \sum_{\mathcal{P}'} \mathbf{Pr}(\mathcal{P}') \sum_{D'_Y} \mathbf{Pr}(D'_Y|\mathcal{P}', D'_X) V(D')$$

We only consider all those D'_X with distinct X_i values. Based on the assumption, such D'_X accounts for at least $1-\delta_2$ of the probability mass. Now the important observation is that for every fixed D'_X with distinct

X values, the marginal distribution of D'_Y given D'_X (i.e. marginalize over $\mathcal{P}'$) is exactly $\mathcal{P}(D'_Y|D'_X)$, the distribution that we sample labels independently from $\mathcal{P}(Y|X)$ for each X'_i in D'_X:

$$\sum_{D'_X}\mathbf{Pr}(D'_X)\sum_{\mathcal{P}'}\mathbf{Pr}(\mathcal{P}')\sum_{D'_Y}\mathbf{Pr}(D'_Y|\mathcal{P}',D'_X)V(D')$$
$$\geq\sum_{D'_X}\mathbf{Pr}(D'_X)\mathbb{1}_{\forall i\neq j,X'_i\neq X'_j}\sum_{D'_Y}\mathbf{Pr}(D'_Y|\mathcal{P},D'_X)V(D')$$

The latter probability is actually the probability that D' will pass V and have distinct X values at the same time. Based on the assumptions in the lemma, it occurs with probability at least $1-\delta_1-\delta_2$. □

Now given this lemma, the proof of Theorem 1 is easy: We show that if any prover A_f satisfies the two conditions in the theorem, it can be used as the verifier V in the lemma such that no $\mathcal{P}'$ can satisfy all three conditions.

Let $\delta_1=\frac{1}{3}$, then the first assumption in the lemma is satisfied, also since $\forall x\in\mathcal{X},\mathcal{Q}(x)<\frac{1}{10n^2}$, we have:

$$\forall i\neq j,\mathbf{Pr}(X_i=X_j)=\sum_x\mathcal{Q}(x)^2\leq\frac{1}{10n^2}$$

By a union bound, we have:

$$\mathbf{Pr}(\forall i\neq j,X_i\neq X_j)\geq\frac{9}{10}$$

Therefore we can set $\delta_2=0.1$. By the above lemma, there exists another $\mathcal{P}'$ such that

$$\mathbf{Pr}_{D'\sim\mathcal{P}'}(A_f(D'))\geq 1-\frac{1}{3}-\frac{1}{10}$$

and

$$\forall X\in\mathcal{X},Y\in\mathcal{Y},\mathcal{P}'(X,Y)=0\ or\ 1$$

On the other hand, note that the l_1 distance between $\mathcal{P}'$ and $\mathcal{P}$ is at least B, then by the properties of A_f, D' cannot pass A_f with probability greater than $\frac{1}{3}$. This contradicts our earlier result. Therefore no such A_f can exist.

□

Proof of Claim 1

Proof. The expected loss is

$$a\mathcal{P}(g(f(X))=1,Y=-1)+b\mathcal{P}(g(f(X))=-1,Y=1)$$

Define $S=\{f(X):X\in\mathcal{X}\}$, then we have:

$$\begin{aligned}&a\mathcal{P}(g(f(X))=1,Y=-1)+b\mathcal{P}(g(f(X))=-1,Y=1)\\=&\sum_{p\in S}\sum_{X:f(X)=p}[a\mathbb{1}_{g(p)=1}\mathcal{P}(Y=-1,X)+\\&b\mathbb{1}_{g(p)=-1}\mathcal{P}(Y=1,X)]\\=&\sum_{p\in S}[a\mathbb{1}_{g(p)=1}\sum_{X:f(X)=p}\mathcal{P}(Y=-1,X)+\\&b\mathbb{1}_{g(p)=-1}\sum_{X:f(X)=p}\mathcal{P}(Y=1,X)]\end{aligned}$$

Therefore, the optimal g^* has $g^*(p)=1$ if and only if:

$$a\sum_{X:f(X)=p}\mathcal{P}(Y=-1,X)\leq b\sum_{X:f(X)=p}\mathcal{P}(Y=1,X)$$

Which is equivalent as:

$$a\mathcal{P}(Y=-1|f(X)=p)\leq b\mathcal{P}(Y=1|f(X)=p)$$

Since f is calibrated, $\mathcal{P}(Y=1|f(X)=p)=p$, therefore $g^*(p)=1$ if and only if $p\geq\frac{a}{a+b}$. □

Proof of Theorem 2

Proof. We will use the following uniform convergence result (Shalev-Shwartz and Ben-David, 2014):

Theorem 3. *Let D be i.i.d. samples of $(\mathcal{X}\times\mathcal{Y},\mathcal{P})$, then with probability at least $1-\delta$,*

$$\begin{aligned}&\sup_{g\in\mathcal{G}}|\tfrac{1}{n}\textstyle\sum_{i=1}^n g(x_i,y_i)-\mathbb{E}g(X,Y)|\\\leq\ &2\mathbb{E}_D R_D(\mathcal{G})+\sqrt{\tfrac{2\ln(4/\delta)}{n}}\end{aligned}\qquad(1)$$

In the following we sometimes allow $\mathcal{G}$ to be a collection of functions from $\mathcal{X}$ to $[0,1]$ in the above results. When used in this sense, we assume that the function will not use y label: $g(x,y)=g(x)$.

Define $\mathcal{F}_{D,p_1,p_2}(f)$ to be the relative frequency of event $\{p_1<f(x)\leq p_2,y=1\}$:

$$\mathcal{F}_{D,p_1,p_2}(f)=\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{p_1<f(x_i)\leq p_2,y_i=1}$$

Define $\mathcal{F}_{\mathcal{P},p_1,p_2}(f)$ to be the probability of the same event:

$$\mathcal{F}_{\mathcal{P},p_1,p_2}(f)=\mathcal{P}(p_1<f(X)\leq p_2,Y=1)$$

Define $\mathcal{E}_{D,p_1,p_2}(f)$ as the empirical expectation of $f(x)\mathbb{1}_{p_1<f(x)\leq p_2}$:

$$\mathcal{E}_{D,p_1,p_2}(f)=\frac{1}{n}\sum_{i=1}^n f(x_i)\mathbb{1}_{p_1<f(x_i)\leq p_2}$$

Define $\mathcal{E}_{\mathcal{P},p_1,p_2}(f)$ as the expectation of the same function:

$$\mathcal{E}_{\mathcal{P},p_1,p_2}(f) = \mathbb{E}[f(X)\mathbb{1}_{p_1<f(X)\le p_2}]$$

When the context is clear, subscripts p_1 and p_2 can be dropped. Using these notations, we can rewrite $c(f)$ and $c_{\text{emp}}(f, D)$ as follows:

$$c(f) = \sup_{p_1,p_2} |\mathcal{F}_{\mathcal{P}}(f) - \mathcal{E}_{\mathcal{P}}(f)|$$

$$c_{\text{emp}}(f) = \sup_{p_1,p_2} |\mathcal{F}_D(f) - \mathcal{E}_D(f)|$$

Note that:

$$\begin{aligned}
&|\sup_{p_1,p_2} |\mathcal{F}_D(f) - \mathcal{E}_D(f)| - \sup_{p_1,p_2} |\mathcal{F}_S(f) - \mathcal{E}_S(f)|| \\
\le &\sup_{p_1,p_2} ||\mathcal{F}_D(f) - \mathcal{E}_D(f)| - |\mathcal{F}_S(f) - \mathcal{E}_S(f)|| \\
\le &\sup_{p_1,p_2} |\mathcal{F}_D(f) - \mathcal{E}_D(f) - \mathcal{F}_S(f) + \mathcal{E}_S(f)| \\
\le &\sup_{p_1,p_2} (|\mathcal{F}_D(f) - \mathcal{F}_S(f)| + |\mathcal{E}_D(f) - \mathcal{E}_S(f)|) \\
\le &\sup_{p_1,p_2} |\mathcal{F}_D(f) - \mathcal{F}_S(f)| + \sup_{p_1,p_2} |\mathcal{E}_D(f) - \mathcal{E}_S(f)|
\end{aligned}$$

Therefore it suffices to show that

$$\begin{aligned}
\mathbf{P}(&\sup_{f,p_1,p_2} |\mathcal{F}_D(f) - \mathcal{F}_S(f)| + \\
&\sup_{f,p_1,p_2} |\mathcal{E}_D(f) - \mathcal{E}_S(f)| > \epsilon) < \delta
\end{aligned}$$

Define

$$\begin{aligned}
\mathcal{H}_1 &= \{\mathbb{1}_{p_1<f(x)\le p_2, y=1} : p_1, p_2 \in \mathbb{R}, f \in \mathcal{F}\} \\
\mathcal{H}_2 &= \{f(x)\mathbb{1}_{p_1<f(x)\le p_2} : p_1, p_2 \in \mathbb{R}, f \in \mathcal{F}\}
\end{aligned}$$

Then we have the following lemma:

Lemma 2. *Let $\mathcal{H}_1, \mathcal{H}_2$ as defined above, then:*

$$R_D(\mathcal{H}_1) \le R_D(\mathcal{H}) \quad R_D(\mathcal{H}_2) \le R_D(\mathcal{H})$$

Proof. For $R_D(\mathcal{H}_1)$, we have:

$$\begin{aligned}
&R_D(\mathcal{H}_1) \\
=&\frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^n \sigma_i \mathbb{1}_{p_1<f(x_i)\le p_2, y_i=1}] \\
=&\frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^n \sigma_i \mathbb{1}_{p_1<f(x_i)\le p_2} \mathbb{E}_{z_i\in\{\pm1\}} \max(z_i, y_i)] \\
\le&\frac{1}{n}\mathbb{E}_{\sigma,z\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^n \mathbb{1}_{p_1<f(x_i)\le p_2} \sigma_i \max(z_i, y_i)] \\
=&\frac{1}{n}\mathbb{E}_{t\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^n t_i \mathbb{1}_{p_1<f(x_i)\le p_2}] \\
=&R_D(\mathcal{H})
\end{aligned}$$

where the last step is because $t_i = \sigma_i \max(z_i, y_i)$ is uniformly distributed over $\{\pm1\}$ independent of the value of y_i.

For $R_D(\mathcal{H}_2)$, we have:

$$\begin{aligned}
&R_{S_n}(\mathcal{H}_2) \\
=&\frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^n \sigma_i f(x_i) \mathbb{1}_{p_1<f(x_i)\le p_2}] \\
=&\frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \int_0^1 \sum_{i=1}^n \sigma_i \mathbb{1}_{t<f(x_i)} \mathbb{1}_{p_1<f(x_i)\le p_2} dt] \\
\le&\frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n} \int_0^1 [\sup_{p_1,p_2,f} \sum_{i=1}^n \sigma_i \mathbb{1}_{\max(p_1,t)<f(x_i)\le p_2}] dt \\
=&\frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n} \int_0^1 [\sup_{p_1'\ge t,p_2,f} \sum_{i=1}^n \sigma_i \mathbb{1}_{p_1'<f(x_i)\le p_2}] dt \\
\le&\frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n} \int_0^1 [\sup_{p_1',p_2,f} \sum_{i=1}^n \sigma_i \mathbb{1}_{p_1'<f(x_i)\le p_2}] dt \\
=&\frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^n \sigma_i \mathbb{1}_{p_1'<f(x_i)\le p_2}] \\
=&R_D(\mathcal{H})
\end{aligned}$$

where the second step is due to $f(x) = \int_0^1 \mathbb{1}_{t<f(x)} dt$, and the forth step is just substituting $\max(p_1, t)$ with p_1'. Since there is no constraint on p_1, the p_1' can take any value greater than or equal to t. □

Combining this lemma with the assumptions in the theorem:

$$\begin{aligned}
\mathbb{E}_D R_D(\mathcal{H}_1) + \sqrt{\frac{2\ln(8/\delta)}{n}} < \frac{\epsilon}{2} \\
\mathbb{E}_D R_D(\mathcal{H}_2) + \sqrt{\frac{2\ln(8/\delta)}{n}} < \frac{\epsilon}{2}
\end{aligned}$$

By Equation (1):

$$\begin{aligned}
\mathbf{P}(\sup_{f,p_1,p_2} |\mathcal{F}_D(f) - \mathcal{F}_S(f)| > \frac{\epsilon}{2}) < \frac{\delta}{2} \\
\mathbf{P}(\sup_{f,p_1,p_2} |\mathcal{E}_D(f) - \mathcal{E}_S(f)| > \frac{\epsilon}{2}) < \frac{\delta}{2}
\end{aligned}$$

□

Proof of Claim 2

Proof. For any $\sigma \in \{\pm1\}^n$, we can find a vector w such that for every X_i, we have $w^T X_i = \sigma_i$ (this is always possible since the number of equations n is less than the dimensionality d). Let $w^* = \frac{Bw}{||w||_2}$ so that $||w^*||_2 = B$, and let $a = \lambda||w||_2/B$ and $b = 0$. Then

1. For $i = 0, \ldots, n$, Compute $P_i = (i, S_i = \sum_{j \le i} \mathbb{1}_{y_j = 1})$
2. Let $cv(P)$ be the convex hull of the set of points P_i
3. For $i = 0, \ldots, n$, Let $Z_i =$ intersection of $cv(P)$ and the line $x = i$
4. Compute $z_i = Z_i - Z_{i-1}$
5. Let $g(f_0(x_i)) = z_i$, extrapolate these points to get continuous nondecreasing function g.

Algorithm 1: Isotonic Regression Calibration Algorithm (PAV Algorithm)

we have:

$$f(X_i) = \frac{1}{1 + \exp(a(w^*)^T x + b)} = \frac{1}{1 + e^{\lambda \sigma_i}}$$

Let $\lambda \to -\infty$, then $\sum_{i=1}^n \sigma_i f(X_i) \to \sum_{i=1}^n \mathbb{1}_{\sigma_i = 1}$, and the conclusion of the claim follows easily. □

The Hypothesis Class $\mathcal{H}$

In Theorem 2, $\mathcal{H}$ is the collection of binary classifiers obtained by thresholding the output of a fuzzy classifier in $\mathcal{F}$. For many hypothesis classes $\mathcal{F}$, the Rademacher Complexity of $\mathcal{H}$ can be naturally bounded. For instance, if $\mathcal{F}$ is the d-dimensional generalized linear classifiers with monotone link function, then $\mathbb{E}_D R_D(\mathcal{H})$ can be bounded by $O(\sqrt{d \log n / n})$. We remark that $\mathcal{H}$ is different from the hypothesis class $\mathcal{H}_{p_1,p_2}$, where the thresholds are fixed in advance:

$$\mathcal{H}_{p_1,p_2} = \{\mathbb{1}_{p_1 < f(X) \le p_2} : f \in \mathcal{H}\}$$

In general, the gap between the Rademacher Complexities of $\mathcal{H}_0$ and $\mathcal{H}_{p_1,p_2}$ can be arbitrarily large. The following example illustrates this point.

Example 2. *Let $\mathcal{X} = \{1, \ldots, n\}$, and $A_1, A_2, \ldots, A_{2^n}$ be a sequence of sets containing all subsets of $\mathcal{X}$. Let $\mathcal{H}$ be the following hypothesis space:*

$$\mathcal{F} = \{f_i(x) = \frac{i}{2^n} - \frac{1}{2^{n+1}} \mathbb{1}_{x \in A_i} : i \in \{1, 2, \ldots, 2^n\}\}$$

Intuitively, $\mathcal{F}$ contains 2^n classifiers, the ith classifier produces a output of either $\frac{i}{2^n}$ or $\frac{i}{2^n} - \frac{1}{2^{n+1}}$ depending on whether $x \in A_i$. One can easily verify that for any p_1, p_2, the VC-dimension (Vapnik and Chervonenkis, 1971) of $\mathcal{H}_{p_1,p_2}$ is at most 2, but the VC-dimension of $\mathcal{H}$ is n.

However, if for any $x \in \mathcal{X}, f \in \mathcal{F}$, we have $f(x) \in P^*$ with $|P^*| < \infty$, then $R_D(\mathcal{H})$ can be bounded using the maximum VC-dimension of $\mathcal{H}_{p_1,p_2}$ and $\log |P^*|$:

Claim 4. *If for any $f \in \mathcal{F}, x \in \mathcal{X}$, we have $f(x) \in P^*$ where P^* is a finite set, and for all $p_1, p_2 \in \mathbb{R}$, the VC-dimension of hypothesis space $\mathcal{H}_{p_1,p_2}$ is at most d, then for any sample D of size n with $n > d + 1$ we have:*

$$R_D(\mathcal{H}) \le \sqrt{\frac{2d(\ln \frac{n}{d} + 1) + 4 \ln(|P^*| + 1)}{n}}$$

Proof. By Massart Lemma (Shalev-Shwartz and Ben-David, 2014), we have:

$$R_D(\mathcal{H}) \le \sqrt{\frac{2 \ln |\mathcal{H}(D)|}{n}}$$

where $\mathcal{H}(D)$ is the restriction of $\mathcal{H}$ to D. It suffices to show that

$$|\mathcal{H}(D)| \le (|P^*| + 1)^2 (en/d)^d$$

Note that

$$\mathcal{H}(D) = \cup_{p_1,p_2} \mathcal{H}_{p_1,p_2}(D)$$

Since $f(x)$ only takes finite possible values, we only need to consider values of p_1, p_2 in $P^* \cup \{-\infty\}$. Therefore by union bound we have

$$|\mathcal{H}(S_n)| \le \sum_{p_1,p_2 \in P^* \cup \{-\infty\}} |\mathcal{H}_{p_1,p_2}(S_n)|$$

Since each $\mathcal{H}_{p_1,p_2}$ has VC-dimension at most d, by Sauer's Lemma (Shalev-Shwartz and Ben-David, 2014):

$$\forall p_1, p_2, |\mathcal{H}_{p_1,p_2}(S_n)| \le (en/d)^d$$

Combining the last two inequalities, we get the desired result. □

Proof of Claim 3

Proof. For reference, the pseudo-code of the PAV algorithm for isotonic regression (Niculescu-Mizil and Caruana, 2005) can be found in Algorithm 1.

Let $z_i = g(f_0(x_i))$, then we can rewrite the objective function as:

$$\max_{a,b} |\sum_{a < i \le b} (\mathbb{1}_{y_i = 1} - z_i)|$$

To prove Algorithm 1 also minimizes this objective function, we first state the minimization problem as a linear programming:

$$\begin{aligned} \min \xi_1 + \xi_2 \quad \textbf{s.t.} \quad & \xi_1, \xi_2 \ge 0 \\ & 0 \le z_1 \le z_2 \le \ldots \le z_n \le 1 \\ & \forall 1 \le k \le n, \sum_{i \le k} z_i \ge \sum_{i \le k} \mathbb{1}_{y_i = 1} - n\xi_1 \\ & \forall 1 \le k \le n, \sum_{i \le k} z_i \le \sum_{i \le k} \mathbb{1}_{y_i = 1} + n\xi_2 \end{aligned}$$

Define $S_k = \sum_{i \le k} \mathbb{1}_{y_i=1}$ and $Z_k = \sum_{i \le k} z_i$. Then we have the following constraints:

$$\forall 1 \le k \le n-1, Z_k - Z_{k-1} \le Z_{k+1} - Z_k$$
$$\forall 1 \le k \le n, S_k - n\xi_1 \le Z_k \le S_k + n\xi_2$$

Let Z_i^* be the solution produced by Algorithm 1, it should be obvious that $Z_i^* \le S_i$ for all i. Therefore,

$$\xi_2^* = \frac{1}{n}\min_i(S_i - Z_i^*) = 0 \quad \xi_1^* = \frac{1}{n}\max_i(S_i - Z_i^*)$$

We need to prove that $\xi_1^* \le \xi_1 + \xi_2$ for every feasible solution (Z_i, ξ_i). Suppose $\xi_1^* = \frac{1}{n}(S_k - Z_i^*)$, and Z_i^* lies on the line segment $\{(j, S_j), (k, S_k)\}$. Then we have:

$$S_i - n\xi_1^* = Z_i^* = \frac{i-j}{k-j}S_k + \frac{k-i}{k-j}S_j$$

Because of the convexity constraint of Z, it must satisfy the following inequality:

$$Z_i \le \frac{i-j}{k-j}Z_k + \frac{k-i}{k-j}Z_j$$

Computing the difference between these two, we get

$$Z_i - S_i + n\xi_1^* \le \frac{i-j}{k-j}(Z_k - S_k) + \frac{k-i}{k-j}(Z_j - S_j)$$

Substituting in

$$Z_i - S_i \ge -n\xi_1 \quad Z_k - S_k \le n\xi_2 \quad Z_j - S_j \le n\xi_2$$

We get

$$n\xi_1^* \le n\xi_1 + n\xi_2$$

which proves the optimality of Z^*. □

Properties of Isotonic Regression

We can prove several interesting properties of isotonic regression using Theorem 2.

Claim 5. *Let g^* be the calibrating function produced by Algorithm 1, then:*

1. *The empirical calibration measure $c_{emp}(g^* \circ f_0, D)$ of the calibrated classifier is always 0.*
2. *For any asymmetric loss $(1-p, p)$ (i.e., each false negative incurs $1-p$ cost and each false positive incurs p cost), the empirical loss of the calibrated classifier is always no greater than the original classifier (both using the optimal decision threshold p):*

$$\sum_{i=1}^{n}[(1-p)\mathbb{1}_{g^*(f_0(x_i)) \le p, y_i=1} + p\mathbb{1}_{g^*(f_0(x_i)) > p, y_i=0}]$$
$$\le \sum_{i=1}^{n}[(1-p)\mathbb{1}_{f_0(x_i) \le p, y_i=1} + p\mathbb{1}_{f_0(x_i) > p, y_i=0}]$$

In particular, when $p = 0.5$, the empirical accuracy of the calibrated classifier is always greater than or equal to the empirical accuracy of the original classifier.

Proof. Throughout the proof, let C be the convex hull computed in Algorithm 1:

$$C = \{(i_0 = 0, 0), (i_1, S_{i_1}), \ldots, (i_{m-1}, S_{i_{m-1}}), (i_m = n, S_n)\}$$

We will use the following notations:

$$z_i = g^*(f_0(x_i)) \quad Z_k = \sum_{i=1}^{k} z_i \quad S_k = \sum_{i=1}^{k} \mathbb{1}_{y_i=1}$$

1. For any p_1, p_2, let l, r be such that:

$$l = \max_{k \le n, z_k \le p_1} k \qquad r = \max_{k \le n, z_k \le p_2} k$$

If no such k exists, let l, r be 0 respectively. By Algorithm 1, we have

$$\forall i_j < k \le i_{j+1}, z_k = \frac{S_{i_{j+1}} - S_{i_j}}{i_{j+1} - i_j}$$

Thus we have $(l, S_l), (r, S_r) \in C$, $Z_l = S_l, Z_r = S_r$, and therefore

$$\sum_{i=1}^{n} \mathbb{1}_{p_1 < z_i \le p_2, y_i=1} - \sum_{i=1}^{n} \mathbb{1}_{p_1 < z_i \le p_2} z_i$$
$$= (Z_r - Z_l) - (S_r - S_l) = 0$$

which implies that $c_{emp}(g^* \circ f_0) = 0$

2. Let $a = \max\{i : f_0(x_i) \le p\}, b = \max\{i : z_i \le p\}$, then we need to show that

$$(1-p)\sum_{i=1}^{b} \mathbb{1}_{y_i=1} + p\sum_{i=b+1}^{n} \mathbb{1}_{y_i=0}$$
$$\le (1-p)\sum_{i=1}^{a} \mathbb{1}_{y_i=1} + p\sum_{i=a+1}^{n} \mathbb{1}_{y_i=0}$$

We consider two separate cases:

(a) $a \le b$, in this case we only need to show that

$$\sum_{i=a+1}^{b} [p\mathbb{1}_{y_i=0} - (1-p)\mathbb{1}_{y_i=1}] \ge 0$$

or equivalently,

$$p[(b-a) - (S_b - S_a)] - (1-p)(S_b - S_a) \ge 0$$

Rearrange terms, it suffices to show

$$p(b-a) - (S_b - S_a) \ge 0$$

Since $S_b = Z_b, S_a \ge Z_a$

$$S_b - S_a \le Z_b - Z_a \le z_b(b-a) \le p(b-a)$$

(b) $a > b$, in this case we only need to show

$$\sum_{i=b+1}^{a} [p\mathbb{1}_{y_i=0} - (1-p)\mathbb{1}_{y_i=1}] \leq 0$$

or equivalently,

$$p[(a-b) - (S_a - S_b)] - (1-p)(S_a - S_b) \leq 0$$

Rearrange terms, it suffices to show

$$p(a-b) - (S_a - S_b) \leq 0$$

Since $S_b = Z_b, S_a \geq Z_a$

$$S_a - S_b \geq Z_a - Z_b \geq z_{b+1}(a-b) \geq p(a-b)$$

□

We can also use Theorem 2 to derive the following non-asymptotic convergence result of Algorithm 1.

Claim 6. *Let $F(t) = \mathcal{P}(f_0(X) \leq t)$ be the distribution function of $f_0(X)$, and define $G(t)$ as:*

$$G(t) = \mathcal{P}(f_0(X) \leq t, Y = 1)$$

Let $cv : [0,1] \to [0,1]$ be the convex hull of all points $(F(t), G(t))$ for all $t \in [0,1]$. Define G_e as:

$$G_e(t) = \mathbb{E}[\mathbb{1}_{f_0(X) \leq t} g^*(f_0(X))]$$

Then under the same condition in Theorem 2,

$$\mathbf{P}(\sup_t |G_e(t) - cv(F(t))| > 2\epsilon) < 5\delta$$

In particular, if $\mathcal{P}(Y = 1|f_0(X))$ is monotonically increasing, then

$$\mathbf{P}(\sup_t |G_e(t) - G(t)| > 2\epsilon) < 5\delta$$

Let us explain the intuition behind this claim: $F(t)$ is the percentage of data points satisfying $f_0(X) \leq t$, and $G(t)$ is $F(t)$ times the conditional probability of $Y = 1$ in the region $\{f_0(X) \leq t\}$. Now consider points $P_i = (i, S_i)$ in Algorithm 1, it is not hard to show that as $n \to \infty$, the limit of points P_i are the curve $(F(t), G(t)), t \in [0,1]$ (after proper scaling). Similarly, $G_e(t)$ is $F(t)$ times the expected value of $g^*(f_0(X))$ in the region $\{f_0(X) \leq t\}$, and it is not hard to show that $(F(t), G_e(t))$ is the limit of (i, Z_i) (after proper scaling). Now the claim states that in the PAV algorithm, $(F(t), G_e(t))$ converge uniformly to the convex hull of $(F(t), G(t))$, which should not be surprising, since we explicitly computed the convex hull of $\{P_i\}$ in Algorithm 1.

When $\mathcal{P}(Y = 1|f_0(X))$ is monotonically increasing w.r.t. $f_0(X)$, $(F(t), G(t))$ is convex, and Claim 6 immediately implies that $G_e(t)$ will converge uniformly to $G(t)$. In this case, the PAV algorithm will eventually recover the "true" link function $g^*(f_0(X)) = \mathcal{P}(Y = 1|f_0(X))$ given sufficient training samples, and Claim 6 provides a rough estimate of the number of samples required to achieve the desired precision.

Proof. Throughout the proof, let C be the convex hull computed in Algorithm 1:

$$C = \{(i_0 = 0, 0), (i_1, S_{i_1}), \ldots, (i_{m-1}, S_{i_{m-1}}), (i_m = n, S_n)\}$$

We will use the following notations:

$$z_i = g^*(f_0(x_i)) \quad Z_k = \sum_{i=1}^{k} z_i \quad S_k = \sum_{i=1}^{k} \mathbb{1}_{y_i=1}$$

We will use the following facts in the proof of Theorem 2:

$$\mathbf{P}(\sup_{g,p_1,p_2} |\mathcal{F}_D(g \circ f_0) - \mathcal{F}_{\mathcal{P}}(g \circ f_0)| > \frac{\epsilon}{2}) < \frac{\delta}{2}$$

$$\mathbf{P}(\sup_{g,p_1,p_2} |\mathcal{E}_D(g \circ f_0) - \mathcal{E}_{\mathcal{P}}(g \circ f_0)| > \frac{\epsilon}{2}) < \frac{\delta}{2}$$

For any $t \in [0,1]$, let g' be any continuous increasing function from $[0,1]$ to $[0,1]$. Let $k = \max\{i : f_0(x_i) \leq t\}, p_1 = -\infty, p_2 = g'(t)$ in the above inequalities, then we have:

$$\mathbf{P}(|\frac{1}{n} S_k - G(t)| > \frac{\epsilon}{2}) < \frac{\delta}{2} \tag{2}$$

$$\mathbf{P}(|\frac{1}{n} \sum_{i=1}^{k} g'(f_0(x_i)) - \mathbb{E}[\mathbb{1}_{f_0(X) \leq t} g'(f_0(X))]| > \frac{\epsilon}{2}) < \frac{\delta}{2}$$

Let g' be such that $||g' - g^*||_\infty < \lambda$, where $\lambda > 0$ can be arbitrarily small. Let $\lambda \downarrow 0$, then the second inequality implies

$$\mathbf{P}(|\frac{1}{n} Z_k - G_e(t)| > \frac{\epsilon}{2}) < \frac{\delta}{2} \tag{3}$$

Let g' be such that $|g'(x) - 1| < \lambda$ for any x. Let $\lambda \downarrow 0$, then the second inequality implies

$$\mathbf{P}(|\frac{1}{n} k - F(t)| > \frac{\epsilon}{2}) < \frac{\delta}{2} \tag{4}$$

For any $t \in [0,1]$, let $k = \max\{i : f_0(x_i) \leq t\}$. Let $[i_{j-1} = l, i_j = r]$ be the segment of C with $l < k \leq r$. Then we have

$$z_{l+1} = \ldots = z_k = \ldots = z_r$$

$$S_l = Z_l = Z_k - (k-l)z_k$$
$$S_r = Z_r = Z_k + (r-k)z_k$$

On the other hand, by (2), with probability at least $1-\delta$:

$$\frac{1}{n}S_l \geq G(f_0(x_l)) - \frac{\epsilon}{2} \quad \frac{1}{n}S_r \geq G(f_0(x_r)) - \frac{\epsilon}{2}$$

Since cv is the convex hull of $(F(t), G(t))$, we have

$$qG(f_0(x_l)) + (1-q)G(f_0(x_r)) \geq cv(F(t))$$

where $q = \frac{F(f_0(x_r))-F(t)}{F(f_0(x_r))-F(f_0(x_l))}$. Combining all, with probability at least $1-\delta$:

$$\frac{1}{n}Z_k + \frac{1}{n}[ql + (1-q)r - k]z_k + \frac{\epsilon}{2} \geq cv(F(t))$$

By (4), with probability at least $1 - \frac{3}{2}\delta$:

$$\frac{1}{n}l \leq F(f_0(x_l)) + \frac{\epsilon}{2} \quad \frac{1}{n}r \leq F(f_0(x_r)) + \frac{\epsilon}{2}$$

$$\frac{1}{n}k \geq F(t) - \frac{\epsilon}{2}$$

Therefore, we have with probability at least $1 - \frac{5}{2}\delta$,

$$\frac{1}{n}Z_k + \frac{3\epsilon}{2} \geq cv(F(t))$$

Then by (3), with probability at least $1 - 3\delta$,

$$G_e(t) + 2\epsilon \geq cv(F(t))$$

Conversely, suppose $(F(t), cv(F(t)))$ is on the line segment between $(F(a), G(a))$ and $(F(b), G(b))$, then

$$G(a) = cv(F(t)) - w(F(t) - F(a))$$
$$G(b) = cv(F(t)) + w(F(b) - F(t))$$

where $w = \frac{G(b)-G(a)}{F(b)-F(a)}$ (if $F(a) = F(b)$ then just let $w = 1$).

By (2) and (3) and the fact that $S_k \geq Z_k$, with probability at least $1 - 2\delta$:

$$G(a) + \epsilon \geq G_e(a) \quad G(b) + \epsilon \geq G_e(b)$$

Also since $(F(t), G_e(t))$ is convex, we have:

$$qG_e(a) + (1-q)G_e(b) \geq G_e(t)$$

where $q = \frac{F(b)-F(t)}{F(b)-F(a)}$. Combining all above, with probability at least $1 - 2\delta$:

$$cv(F(t)) + \epsilon \geq G_e(t)$$

Combining two directions, the proof is complete. □

Discussion on Kakade's Algorithm (2011)

Kakade's algorithm minimizes the following squared loss objective function:

$$\mathcal{L}(u, w) = \sum_{i=1}^{n}(y_i - u(w \cdot x_i))$$

where u is a non-decreasing 1-Lipschitz function and w satisfies $||w|| \leq W$. In each iteration, the algorithm first fix u and search for the optimal w that minimizes the squared loss, then fix w and run a slightly modified version of the PAV algorithm (Algorithm 1) to find the optimal u.

In Claim 5, we proved that the PAV algorithm always produce a calibrated classifier, therefore Kakade's algorithm can be viewed as alternating between the following two steps:

1. Search for the parameter w that minimizes the squared loss $\mathcal{L}(u, w)$.
2. Find the link function u such that $u(w \cdot x)$ is empirically calibrated.

In other words, each iteration of Kakade's algorithm can be viewed as first optimizing the objective function $\mathcal{L}(u, w)$, then projecting $u(w \cdot x)$ onto the space of empirically calibrated classifiers. An interesting question here is whether the algorithm would still work if we replace the squared loss function with any other loss function in the first step.