Example of Franke Wolfe (for 3 examples)

Example 1: Advice1 applies and target is preferred, hence gradient = 1 – P. Similarly, Advice2 gradient is 1 - P
Example2: 1 – P, 0 – P (Advice2 applies and target is non preferred)
Example3: 0 – P, 0 – P

Data gradient for all three examples is I – P, which is lets say 1 – P, 0 – P, 0 – P.

Linear program now is maximize:

$$\lambda_1 * |1 - P| + \lambda_2 * |1 - P| + \lambda_3 * |1 - P| \ (for\ example\ 1)$$

Similarly for example2 and example3, (a modulus is needed because otherwise it will always prefer positive gradients.

Subject to: $0 \leq \lambda_i \leq 1$ and $\sum \lambda_i = 1$ for each example.

The lambda's give a ranking over which human contributed more to the gradient.

If instead of modulus, the sign operator is used and all the lambda's for advice set to alpha, we get advice gradient $\alpha * (n_t - n_f)$. Therefore, Philip's framework is a subset of this approach.

Problems?

If there are 50 examples, there are 50 equations and constraints!

Why not explore use of **aggregators over mini batches of examples**?

For example for advice 1, $\sum_{ex} I_{a1} - P$ where I_{a1} is 1 or 0 if advice applies and target is preferred or non preferred. This is an example of a mini batch gradient with the sum aggregator.

We could explore use of **max, or weighted sum** if some examples are deemed more important than others for class imbalance.

Ethical concerns?

What if there are 5 people giving advice and the weight on the data gradient still overrides them? This may sometimes be preferred. (Copernicus said heliocentric based on data against all expert advice on geocentric model at the time). But lets say it is a breach of ethics.

For this reason let's define a ethical constant k.

If k > some number of people, then we solve the linear program only over advice gradients and then project the data gradients onto the advice gradients by vector projection.

Consider for 2 examples:
Example1: Data gradient = 1 – P , Advice1 gradient = 1 – P and Advice2 gradient = 0 – P
Let's say on solving the linear program for only the advice gradients, $\lambda_1 = 1 \; and \; \lambda_2 = 0$ and therefore total advice gradient is 1 – P
Similarly for example2: Data gradient is 0 – P and advice gradient was also 0 – P

Now we have data gradient vector v = [1 – P, 0 – P] and advice gradient vector u = [1 – P, 0 – P], so when the ethical constant is exceeded, we can compute the gradient as vector projection of v on u for each example to conform with the advice when the number of humans commenting on the example is too many for the data to hold precedence.