# Statistics 221 Final Project: C++ team

Adam Petcher, Brandon Sim, Andrew Liu, Jason Sclar, Rohit Ramani

December 17, 2013

## 1 Compare MLE to SGD

MLE methods such Newton-Raphson compute the gradient of the log-likelihood, $\nabla \ell$ for all observations in one step. While this works well for small samples, it is too computationally expensive to be done with especially large amounts of data. Sakrison's method relies on the assumption that the expectation of the gradient for a single observation is proportional to the gradient for all observations. By the law of large numbers, repeatedly updating using the gradient for single observations will converge to expectations. This makes it possible to do the update step for a single observation at a time and compute the MLE.

The implicit updates use the update step, $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + a_t \nabla \ell(\boldsymbol{\theta}_{t+1}; y_t, \boldsymbol{x}_t)$ instead of $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + a_t \nabla \ell(\boldsymbol{\theta}_t; y_t, \boldsymbol{x}_t)$. However the expectation of the gradient is the same for all observations. Therefore $\nabla \ell(\boldsymbol{\theta}_{t+1}; y_t, \boldsymbol{x}_t) = \nabla \ell(\boldsymbol{\theta}_t; y_t, \boldsymbol{x}_t)$. Since the implicit method has the same expectation as Sakrison's method, it must also compute the MLE.

## 2 GLM Proofs

(a) Show $E(y_t|\boldsymbol{x}_t) = h(\boldsymbol{x}_t^T \boldsymbol{\theta}^*) = b'(\eta_t)$

We start with the moment generating function, which we solve for using LOTUS

$$
\begin{aligned}
M_Y(t) &= E[e^{tY}] \\
&= \int_Y \exp(ty_k) f(y_k|\eta_k) dy_k \\
&= \int_Y \exp(ty_k) \exp\left(\frac{\eta_t y_t - b(\eta_t)}{\phi}\right) \cdot c(y_t, \phi) dy_k \\
&= \int_Y c(y_t, \phi) \exp\left(ty_k + \frac{\eta_t y_t}{\phi} - \frac{b(\eta_t)}{\phi}\right) dy_k \\
&= \int_Y c(y_t, \phi) \exp\left(\left(\frac{\eta_t y_t + \phi t}{\phi}\right) y_k - \frac{b(\eta_t)}{\phi}\right) dy_k \\
&= \int_Y c(y_t, \phi) \exp\left(\left(\frac{\eta_t y_t + \phi t}{\phi}\right) y_k - \frac{b(\eta_k + \phi t)}{\phi} + \frac{b(\eta_k + \phi t)}{\phi} - \frac{b(\eta_t)}{\phi}\right) dy_k \\
&= \int_Y \exp\left(\frac{b(\eta_k + \phi t) - b(\eta_t)}{\phi}\right) c(y_t, \phi) \exp\left(\frac{(\eta_t y_t + \phi t) y_k - b(\eta_k + \phi t)}{\phi}\right) dy_k \\
&= \exp\left(\frac{b(\eta_k + \phi t) - b(\eta_t)}{\phi}\right) \int_Y c(y_t, \phi) \exp\left(\frac{(\eta_t y_t + \phi t) y_k - b(\eta_k + \phi t)}{\phi}\right) dy_k
\end{aligned}
$$

The remaining integral is the conditional pdf of $y|\eta_k + \phi t$ and integrates to 1, which leaves the remaining MGF for $Y$

$$
M_Y(t) = \exp\left(\frac{b(\eta_k + \phi t) - b(\eta_t)}{\phi}\right)
$$

We then use the MGF to find $E(y_t|\boldsymbol{x}_t)$. We start by finding $M'_Y(t)$

$$\begin{aligned} M'_Y(t) &= \frac{\partial}{\partial t}\left[\exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)\right] \\ &= \exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)\cdot\frac{b'(\eta_k+\phi t)\phi}{\phi} \\ &= \exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)\cdot b'(\eta_k+\phi t) \end{aligned}$$

We then evaluate at $t=0$ which is the first moment.

$$\begin{aligned} E(y_t|\boldsymbol{x}_t) &= M'_Y(0) \\ &= \exp\left(\frac{b(\eta_k+\phi\cdot 0)-b(\eta_t)}{\phi}\right)\cdot b'(\eta_k+\phi\cdot 0) \\ &= \exp\left(\frac{0}{\phi}\right)\cdot b'(\eta_k) \\ &= b'(\eta_k) \end{aligned}$$

(b) Show $\text{Var}(y_t|\eta_t)=\phi\cdot h'(\eta_t)$
We find the variance using the moment generating function from part (a).

$$\text{Var}(y_k|\eta_k) = E(y_k^2|\eta_k) - [E(y_k|\eta_k)]^2$$

We can use the value of $E(y_k|\eta_k)$ from part (a). We then find the second moment $E(y_k^2|\eta_k)$ with our MGF.

$$\begin{aligned} M''_Y(t) =&\frac{\partial}{\partial t}\left[\exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)\cdot b'(\eta_k+\phi t)\right] \\ =&\exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)\cdot b'(\eta_k+\phi t)\cdot b'(\eta_k+\phi t)+ \\ &b''(\eta_k+\phi t)\cdot\phi\cdot\exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right) \\ =&\exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right)\cdot\left[b'(\eta_k+\phi t)\right]^2+b''(\eta_k+\phi t)\cdot\phi\cdot\exp\left(\frac{b(\eta_k+\phi t)-b(\eta_t)}{\phi}\right) \end{aligned}$$

Evaluating at 0 gives

$$\begin{aligned} E(y_k^2|\eta_k) =&M''_Y(0) \\ =&\exp\left(\frac{b(\eta_k+\phi*0)-b(\eta_t)}{\phi}\right)\left[b'(\eta_k+\phi*0)\right]^2+b''(\eta_k+\phi*0)\phi\exp\left(\frac{b(\eta_k+\phi*0)-b(\eta_t)}{\phi}\right) \\ =&\exp(\frac{0}{\phi})\left[b'(\eta_k)\right]^2+\phi\cdot b''(\eta_k)\exp(\frac{0}{\phi}) \\ =&\left[b'(\eta_k)\right]^2+\phi\cdot b''(\eta_k) \end{aligned}$$

Putting the two parts together gives the solution

$$\begin{aligned}\mathrm{Var}(y_t|\eta_t) &= E(y_k^2|\eta_k) - [E(y_k|\eta_k)]^2 \\ &= [b'(\eta_k)]^2 + \phi \cdot b''(\eta_k) - [b'(\eta_k)]^2 \\ &= \phi \cdot b''(\eta_k)\end{aligned}$$

From part (a) we know that $b'(\eta_t) = h(\eta)$ therefore $b''(\eta_t) = h'(\eta_t)$ which gives

$$\mathrm{Var}(y_t|\eta_t) = \phi \cdot h'(\eta_t)$$

(c) Show $\nabla \ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t) = \frac{1}{\phi}(y_t - h(\boldsymbol{x}_t^T \theta))\boldsymbol{x}_t$
To find $\nabla \ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t)$, we take the partial derivative with respect to $\boldsymbol{\theta}$. The likelihood is proportional to the pdf.

$$\begin{aligned}L(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t) &\propto \exp\left(\frac{\eta_t y_t - b(\eta_t)}{\phi}\right) \cdot c(y_t, \phi) \\ \ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t) &= \frac{\eta_t y_t - b(\eta_t)}{\phi} \\ \nabla \ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t) &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[\ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t)\right] \\ &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[\frac{\eta_t y_t - b(\eta_t)}{\phi}\right] \\ &= \frac{\partial}{\partial \boldsymbol{\theta}}\left[\frac{\boldsymbol{x}_t^T \boldsymbol{\theta} y_t - b(\boldsymbol{x}_t^T \boldsymbol{\theta})}{\phi}\right] \\ &= \frac{1}{\phi}(\boldsymbol{x}_t y_t - \boldsymbol{x}_t b'(\boldsymbol{x}_t^T \boldsymbol{\theta})) \\ &= \frac{1}{\phi}(y_t - h(\boldsymbol{x}_t^T \boldsymbol{\theta}))\boldsymbol{x}_t\end{aligned}$$

(d) Show $\mathcal{J}(\boldsymbol{\theta}) = -E(\nabla\nabla \ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t)) = \frac{1}{\phi} E(h'(\boldsymbol{x}_t^T \boldsymbol{\theta})\boldsymbol{x}_t \boldsymbol{x}_t')$
To find the Fisher information, we take the negative expectation of the second partial derivative of the log-likelihood with respect to $\boldsymbol{\theta}$. We start with the the solution from part (c).

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}) &= -E\Bigg(\nabla\nabla \ell(\boldsymbol{\theta}; y_t, \boldsymbol{x}_t)\Bigg) \\ &= -E\Bigg(\frac{\partial}{\partial \boldsymbol{\theta}}\left[\frac{1}{\phi}(y_t - h(\boldsymbol{x}_t^T \boldsymbol{\theta}))\boldsymbol{x}_t\right]\Bigg) \\ &= -E\Bigg(-\frac{1}{\phi} \cdot h'(\boldsymbol{x}_t^T \boldsymbol{\theta})\boldsymbol{x}_t \boldsymbol{x}_t^T\Bigg) \\ &= \frac{1}{\phi}\Bigg(h'(\boldsymbol{x}_t^T \boldsymbol{\theta})\boldsymbol{x}_t \boldsymbol{x}_t^T\Bigg)\end{aligned}$$

## 3 Implementation and Results

(a) **Log-loss.** The log-loss function is given by:

$$L(y, \hat{y}) = \log(1 + \exp(-y\hat{y})) \tag{1}$$

The SGD update is given by:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) \tag{2}$$

where

$$Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = \log\left(1 + \exp(-y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t})\right) + \frac{\lambda}{2}\|\boldsymbol{\theta_t}\|^2 \tag{3}$$

The gradient $\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t})$ can be calculated as follows:

$$\begin{aligned}
\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) &= \left(\frac{1}{1 + \exp(-y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t})}\right)\left(\exp(-y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t})\right)(-y_t \cdot \boldsymbol{x_t}) + \lambda\boldsymbol{\theta_t} \\
&= \frac{-y_t \exp(-y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t})}{1 + \exp(-y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t})} \cdot \boldsymbol{x_t} + \lambda\boldsymbol{\theta_t}
\end{aligned}$$

The implicit update can be derived in much the same way as above. We have the following:

$$\begin{aligned}
&= \frac{-y_t}{\exp(y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t}) + 1} \cdot \boldsymbol{x_t} + \lambda\boldsymbol{\theta_t}\boldsymbol{\theta_{t+1}} \qquad &= \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) \\
&= \boldsymbol{\theta_t} - \alpha_t\left(\frac{-y_t}{\exp(y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_{t+1}}) + 1} \cdot \boldsymbol{x_t} + \lambda\boldsymbol{\theta_{t+1}}\right)
\end{aligned}$$

**Hinge-loss.** The hinge loss function is given by:

$$L(y, \hat{y}) = \max(0, 1 - y\hat{y}) \tag{4}$$

The SGD update is given by:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) \tag{5}$$

where

$$Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = \max(0, 1 - y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t}) + \frac{\lambda}{2}\|\boldsymbol{\theta_t}\|^2 \tag{6}$$

The gradient $\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t})$ can be calculated as follows, where we consider two cases depending on the sign of $1 - y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t}$:

1. If $1 - y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t} < 0$, then $Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = \frac{\lambda}{2}\|\boldsymbol{\theta_t}\|^2$. Then,

$$\begin{aligned}
\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) &= \nabla\left(\frac{\lambda}{2}\|\boldsymbol{\theta_t}\|^2\right) \\
&= \lambda\boldsymbol{\theta_t}
\end{aligned}$$

2. If $1 - y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t} \geq 0$, then $Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = 1 - y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t} + \frac{\lambda}{2}\|\boldsymbol{\theta_t}\|^2$. Then,

$$\begin{aligned}
\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) &= \nabla\left(1 - y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t} + \frac{\lambda}{2}\|\boldsymbol{\theta_t}\|^2\right) \\
&= \nabla\left(-y_t \cdot \boldsymbol{x_t^T}\boldsymbol{\theta_t}\right) + \lambda\boldsymbol{\theta_t} \\
&= -y_t \cdot \boldsymbol{x_t} + \lambda\boldsymbol{\theta_t}
\end{aligned}$$

where the last step is accomplished by noting that:

$$\begin{aligned}
\nabla_\theta \left(\boldsymbol{x^T\theta}\right) &= \nabla_\theta \left(\sum_{i=1}^{|\boldsymbol{x}|} x_i\theta_i\right) \\
&= \left(\frac{\partial}{\partial\theta_1}\left(\sum_{i=1}^{|\boldsymbol{x}|} x_i\theta_i\right), \frac{\partial}{\partial\theta_2}\left(\sum_{i=1}^{|\boldsymbol{x}|} x_i\theta_i\right), \ldots\right) \\
&= (x_1, x_2, \ldots) \\
&= \boldsymbol{x}
\end{aligned}$$

Putting these results together, we have that:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \begin{cases} \lambda\boldsymbol{\theta_t} & y_t \cdot \boldsymbol{x_t^T\theta_t} > 1 \\ \lambda\boldsymbol{\theta_t} - y_t\boldsymbol{x_t} & \text{otherwise} \end{cases} \tag{7}$$

We now derive the implicit update for the hinge loss, which is given by:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) \tag{8}$$

The calculations are very similar to the SGD derivation above, and for the two cases depending on the sign of $1 - y_t \cdot \boldsymbol{x_t^T\theta_{t+1}}$ are:

1. If $1 - y_t \cdot \boldsymbol{x_t^T\theta_{t+1}} < 0$, then

$$\nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) = \lambda\boldsymbol{\theta_{t+1}}$$

Then, substituting this result into the implicit update equation above, we can solve to find:

$$\boldsymbol{\theta_{t+1}} = \frac{1}{1+\lambda\alpha_t}\boldsymbol{\theta_t} \tag{9}$$

2. If $1 - y_t \cdot \boldsymbol{x_t^T\theta_{t+1}} \geq 0$, then

$$\nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) = -y_t \cdot \boldsymbol{x_t} + \lambda\boldsymbol{\theta_{t+1}}$$

Then, substituting this result into the implicit update equation above, we can solve to find:

$$\boldsymbol{\theta_{t+1}} = \frac{1}{1+\lambda\alpha_t}\left(\boldsymbol{\theta_t} + \alpha_t y_t \cdot \boldsymbol{x_t}\right) \tag{10}$$

Note that during implementation, we should be careful to make sure that we check the sign of $1 - y_t \cdot \boldsymbol{x_t^T\theta_{t+1}}$ after the update, as the derivation of the implicit updates forced an assumption of the sign to begin with. If the assumption was wrong, then the other update function should be used.

**Squared-loss.** The squared-loss function is given by:

$$L(y, \hat{y}) = (y - \hat{y})^2 \tag{11}$$

The SGD update is given by:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} - \alpha_t \nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) \tag{12}$$

where

$$Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) = \left(y_t - \boldsymbol{x_t^T}\boldsymbol{\theta_t}\right)^2 + \frac{\lambda}{2}\|\boldsymbol{\theta_t}\|^2 \tag{13}$$

The gradient $\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t})$ can be calculated as follows:

$$\begin{aligned}\nabla Q(\boldsymbol{\theta_t}, y_t, \boldsymbol{x_t}) &= 2\left(y_t - \boldsymbol{x_t^T}\boldsymbol{\theta_t}\right)(-\boldsymbol{x_t}) + \lambda\boldsymbol{\theta_t} \\ &= -2(y_t - \boldsymbol{x_t^T}\boldsymbol{\theta_t})\boldsymbol{x_t} + \lambda\boldsymbol{\theta_t}\end{aligned}$$

So, the SGD update for the log-loss function is:

$$\boldsymbol{\theta_{t+1}} = \boldsymbol{\theta_t} + 2\alpha_t\left(y_t - \boldsymbol{x_t^T}\boldsymbol{\theta_t}\right)\boldsymbol{x_t} - \alpha_t\lambda\boldsymbol{\theta_t} \tag{14}$$

The implicit update can be derived in much the same way as above. We have the following:

$$\begin{aligned}\boldsymbol{\theta_{t+1}} &= \boldsymbol{\theta_t} - \alpha_t\nabla Q(\boldsymbol{\theta_{t+1}}, y_t, \boldsymbol{x_t}) \\ \boldsymbol{\theta_{t+1}} &= \boldsymbol{\theta_t} + 2\alpha_t\left(y_t - \boldsymbol{x_t^T}\boldsymbol{\theta_{t+1}}\right)\boldsymbol{x_t} - \alpha_t\lambda\boldsymbol{\theta_{t+1}} \\ (1+\alpha_t\lambda)\boldsymbol{\theta_{t+1}} &= \boldsymbol{\theta_t} + 2\alpha_t\left(y_t - \boldsymbol{x_t^T}\boldsymbol{\theta_{t+1}}\right)\boldsymbol{x_t} \\ (1+\alpha_t\lambda)\boldsymbol{\theta_{t+1}} &= \boldsymbol{\theta_t} + 2\alpha_t y_t\boldsymbol{x_t} - 2\alpha_t\boldsymbol{x_t^T}\boldsymbol{\theta_{t+1}}\boldsymbol{x_t} \\ \left[(1+\alpha_t\lambda)I + 2\alpha_t\boldsymbol{x_t^T}\boldsymbol{x_t}I\right]\boldsymbol{\theta_{t+1}} &= \boldsymbol{\theta_t} + 2\alpha_t y_t\boldsymbol{x_t} \\ \boldsymbol{\theta_{t+1}} &= \left[(1+\alpha_t\lambda)I + 2\alpha_t\boldsymbol{x_t^T}\boldsymbol{x_t}I\right]^{-1}(\boldsymbol{\theta_t} + 2\alpha_t y_t\boldsymbol{x_t})\end{aligned}$$

(b) We downloaded and compiled Bottou's SGD package.

(c) See `svmimplicit.cpp`.

(d) Here we report the results from running the SGD experiment. We compare performance of baseline SGD, ASGD, and our implicit SGD for each of the log loss and hinge loss functions. (We leave out LibLinear because we could not successfully run the training program without using too much memory.) Unless otherwise stated, SGD algorithms use their default regularization parameters (close to $\lambda = 10^{-5}$) and learning rate $\eta_t = \frac{\eta_0}{1+\lambda\eta_0 t}$, with $\eta_0$ picked by the Bottou package unless otherwise specified. We measure cost as

$$C = L + 0.5\lambda||\Theta||^2$$

where $L$ denotes the average loss over all test runs and $\Theta$ the parameter vector.

1. RCV1 Benchmark, hinge-loss.

| algorithm | training time (s) | test error (%) | cost |
|---|---|---|---|
| SGD | 0.87 | 6.005 | 0.244 |
| ASGD | 0.86 | 6.018 | 0.244 |
| Implicit | 0.71 | 5.184 | 0.144 |

2. RCV1 Benchmark, log-loss.

| algorithm | training time (s) | test error (%) | cost |
|---|---|---|---|
| SGD | 3.47 | 5.175 | 0.153 |
| ASGD | 4.03 | 5.145 | 0.154 |
| Implicit ($\eta_0 = 10$, not 4) | 1.92 | 5.262 | 0.537 |

For implicit SGD, I use learning rate $\eta_t = \frac{\eta_0}{1+\lambda\eta_0 t/2}$.

3. Alpha dataset, hinge-loss.

| algorithm | training time (s) | test error (%) | cost |
|---|---|---|---|
| SGD | 3.37 | 22.7 | 0.548 |
| ASGD | 2.27 | 21.83 | 0.532 |
| Implicit ($\eta_0 = 10$, not 0.25) | 2.29 | 22.18 | 0.565 |

Implicit SGD here was improved by using default $\lambda = 10^{-5}$ (whereas SGD and ASGD used $\lambda = 10^{-6}$).

4. Alpha dataset, log-loss.

| algorithm | training time (s) | test error (%) | cost |
|---|---|---|---|
| SGD | 9.31 | 22.11 | 0.477 |
| ASGD | 3.5 | 21.9 | 0.474 |
| Implicit ($\eta_0 = 10$, not 0.25) | 3.7 | 21.9 | 0.542 |

# 4 Contributions

On the implementation side, Adam did a majority of the work, being the first to download Bottou's SGD package and get it working (1.4b), and implementing each of the loss functions with implicit update (1.4c), as well as working with Brandon to figure out some of the math associated with those updates. He also debugged the `svmimplicit.cpp` code and did a little tuning of the parameters.

On the analysis side, Brandon derived the SGD and implicit updates for the three loss functions (1.4a). Jason gave the intuition for Sakrison's method and the implicit method computing the MLE (1.1), and performed the derivations of the desired properties of GLMs (1.2).

Finishing off the implementation, Andrew ran Adam's implicit SGD, along with baseline SGD and AGSD, on RCV1 and alpha datasets, collecting training time, test error, and cost data (1.4d, e). He tried but failed to get LibLinear working. He also did a little work to compile each team member's tex into this writeup.

Adam and Brandon made large contributions in terms of keeping the team going forward, communicating their progress to the group and ensuring the group stayed knowledgeable, and starting on their work early so that the rest of the team had an easier time finishing off the missing pieces at the end.