$Y_i \sim \text{Poisson} \left( \mu_i \right)$

$\mu_i = e^{x_i \cdot w}$

$x_i, w \in \mathbb{R}^d$

$\forall i \neq j, \ Y_i \perp Y_j \mid \mu_i, \mu_j$

This is kinda like logistic regression. Both are special cases of the generalized linear model.

Well, "accelerated," at least...

Suppose $d > m$. We will implement a fast algorithm to minimize:

$$f(w) = -L(w) + \lambda \|w\|_1$$

"$\lambda$-regularized Poisson regression"

where $L(w)$ is the likelihood function. Todo:

0. Write equivalent objective $g(w) = L(w) + \lambda \|w\|_1$.

1. Recall accelerated gradient method, requirements differentiable convex.

2. Recall closed-form solution to prox function.

3. Compute $\partial L(w)/\partial w$ using matrix differentials.

We'll finish up with some matrix differentials.

(Not in these notes, though.)
Since the $y_i$ are (conditionally) independent,

$$L(w) = \prod_{i=1}^{m} p(y_i | x_i, w)$$

$$= \prod_{i=1}^{m} \frac{\exp(y_i(x_i \cdot w)) \exp[-\exp(x_i \cdot w)]}{y_i!}$$

$$\log L(w) = \sum_{i=1}^{m} (y_i(x_i \cdot w) - \exp(x_i \cdot w)) - \sum_{i=1}^{m} y_i!$$

$$-\lambda(w)$$

Equivalently, minimize $g(w)$ as previously defined.

1. Accelerated gradient method:

$$w^{(-1)} = w^{(0)} = 0$$

$$v = w^{(k-1)} + \frac{k-2}{k+1} (w^{(k-1)} - w^{(k-2)})$$

$$w^{(k)} = \text{prox}_{t_k} \left( v - t_k \frac{\partial \lambda(v)}{\partial v} \right)$$

$$\text{prox}_{t_k}(u) = \arg \min_{z \in \mathbb{R}^d} \frac{1}{2t_k} \| z - u \|^2 + \gamma \| u \|,$$

2. $S_{\gamma t_k}(u)$

$$= \left\{ \begin{array}{ll}
    u_i + \gamma t_k & \text{if } u_i < \gamma t_k \\
    0 & \text{if } -\gamma t_k \leq u_i \leq \gamma t_k \\
    u_i - \gamma t_k & \text{if } u_i > \gamma t_k
\end{array} \right\}_{1 \leq i \leq d}$$
Now in matrix form:

\[
X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad Xw = \begin{bmatrix} x_1 w \\ \vdots \\ x_n w \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & y_n \end{bmatrix}, \quad YXw = \begin{bmatrix} y_1(x_1w) \\ \vdots \\ y_n(x_nw) \end{bmatrix}
\]

\[
\ell(w) = 1^T(\exp(Xw) - YXw)
\]

\[
d\ell(w) = 1^T(d\exp(Xw) - YX\,dw)
\]

\[
\begin{align*}
&d\exp(w) = \exp(w) \cdot dw \\
&dXw = X\,dw \\
&d\exp(Xw) = \exp(Xw) \cdot X\,dw
\end{align*}
\]

\[
= 1^T(\exp(Xw)X\,dw - YX\,dw)
\]

Thus

\[
W^{(k)} = S_{\lambda t_k} \left[ V + \left( Y - \text{diag}(\exp(Xv)) \right) X \, t_k \right]
\]

Recall \( y_i \sim \text{Poisson} \left( \exp\left( x_i \cdot w \right) \right) \), so a perfect fit makes sense. (Of course, we still want to "shrink" it to regularize.)