CMU 10-725 F12

SHIVA KAUL ⟨SKKAUL@CS.CMU.EDU⟩

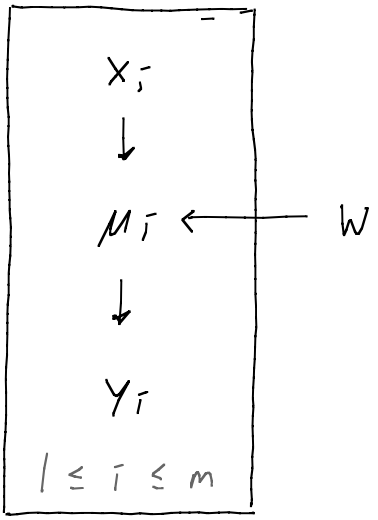
RECITATION 5: POISSON REGRESSION, ACCELERATED GRADIENT, AND DIFFERENTIALS

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

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

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

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



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

Suppose $d > m$. We will implement a fast (well, "accelerated," at least...) algorithm to minimize:

$$f(w) = -L(w) + \lambda \|w\|_1, \qquad \text{"}\ell_1\text{-regularized Poisson regression"}$$

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

0. write equivalent objective $g(w) = \ell(w) + \gamma\|w\|_1$ (ℓ: differentiable, convex)
1. recall accelerated gradient method, requirements
2. recall closed-form solution to prox function
3. compute $\partial \ell(w) / dw$ using matrix differentials

We'll finish up with some matrix differentials.
(Not in these notes, though.)

0 since the y_i are (conditionally) independent,

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

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

$$p(y) = \frac{\mu^y}{y!} \cdot e^{-\mu}$$

POISSON PDF

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

irrelevant to optimization

equivalently minimize $g(w)$ as previously defined.

1 accelerated gradient method:

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

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

$$w^{(k)} = \text{prox}_{t_k}\left(v - t_k \, \partial \ell(v) / \partial v\right)$$

$$\text{prox}_{t_k}(u) = \underset{z \in \mathbb{R}^d}{\text{argmin}} \; \frac{1}{2t_k} \|z - u\|^2 + \gamma \|u\|_1$$

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

$$= \left[\begin{cases} u_i + \gamma t_k & \text{if } u_i < \gamma t_k \\ 0 & \text{if } -\gamma t_k \le u_i \le \gamma t_k \\ u_i - \gamma t_k & \text{if } u_i > \gamma t_k \end{cases} \right] \quad 1 \le i \le d$$

3 Now in matrix form:

$$X = \begin{bmatrix} — & x_1 & — \\ & \vdots & \\ — & x_n & — \end{bmatrix}, \quad Xw = \begin{bmatrix} x_1 \cdot w \\ \vdots \\ x_n \cdot w \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 & & 0 \\ & \ddots & \\ 0 & & y_n \end{bmatrix}, \quad YXw = \begin{bmatrix} y_1 (x_1 \cdot w) \\ \vdots \\ y_n (x_n \cdot w) \end{bmatrix}$$

$$\ell(w) = 1^T \left(\exp(Xw) - YXw \right)$$

$$d\ell(w) = 1^T \left(d\exp(Xw) - YX\,dw \right)$$

$$\begin{aligned} d\exp(w) &= \exp(w) \circ dw \\ dXw &= X\,dw \\ d\exp(Xw) &= \exp(Xw) \circ X\,dw \end{aligned}$$

not the most thrilling application of matrix differentials...

$$= 1^T \left(\exp(Xw) \circ X\,\partial w - YX\,\partial w \right)$$

Thus $$w^{(k)} = S_{\lambda t_k} \left[v + \left(y - \operatorname{diag}(\exp(Xv)) \right) X t_k \right]$$

Recall $y_i \sim \text{Poisson}(\exp(x_i \cdot w))$, so a perfect fit

$$Y = \begin{bmatrix} \exp(x_1 \cdot v) \\ \vdots \\ \exp(x_n \cdot v) \end{bmatrix}$$

makes sense. (Of course, we still want to "shrink" it to regularize.)