Conditional Probability Models

David Rosenberg

New York University

October 29, 2016

Maximum Likelihood Estimation

Estimating a Probability Distribution: Setting

- Let $p(y)$ represent a probability distribution on $\mathcal{Y}$.
- $p(y)$ is **unknown** and we want to **estimate** it.
- Assume that $p(y)$ is either a
  - probability density function on a continuous space $\mathcal{Y}$, or a
  - probability mass function on a discrete space $\mathcal{Y}$.
- Typical $\mathcal{Y}$'s:
  - $\mathcal{Y} = \mathbf{R}$; $\mathcal{Y} = \mathbf{R}^d$ [typical continuous distributions]
  - $\mathcal{Y} = \{-1, 1\}$ [e.g. binary classification]
  - $\mathcal{Y} = \{0, 1, 2, \ldots, K\}$ [e.g. multiclass problem]
  - $\mathcal{Y} = \{0, 1, 2, 3, 4 \ldots\}$ [unbounded counts]

Evaluating a Probability Distribution Estimate

- Before we talk about estimation, let's talk about evaluation.
- Somebody gives us an estimate of the probability distribution

$$\hat{p}(y).$$

- How can we evaluate how good it is?
- We want $\hat{p}(y)$ to be descriptive of **future** data.

Likelihood of a Predicted Distribution

- Suppose we have

$$\mathcal{D} = \{y_1, \ldots, y_n\} \text{ sampled i.i.d. from } p(y).$$

- Then the **likelihood** of $\hat{p}$ for the data $\mathcal{D}$ is defined to be

$$\hat{p}(\mathcal{D}) = \prod_{i=1}^{n} \hat{p}(y_i).$$

- We'll write this as

$$L_{\mathcal{D}}(\hat{p}) := \hat{p}(\mathcal{D})$$

- Special case: If $\hat{p}$ is a probability mass function, then
  - $L_{\mathcal{D}}(\hat{p})$ is the probability of $\mathcal{D}$ under $\hat{p}$.

Parametric Models

Definition

A **parametric model** is a set of probability distributions indexed by a parameter $\theta \in \Theta$. We denote this as

$$\{p(y;\theta) \mid \theta \in \Theta\},$$

where θ is the **parameter** and Θ is the **parameter space**.

- In **probabilistic modeling,** analysis begins with something like:

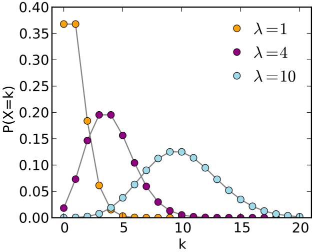
  Suppose the data are generated by a distribution in parametric family $\mathcal{F}$ (e.g. a Poisson family).

- Our perspective is different, at least conceptually:
  - We don't make any assumptions about the data generating distribution.
  - We use a parametric model as a **hypothesis space**.
  - (More on this later.)

Poisson Family

- Support $\mathcal{Y} = \{0, 1, 2, 3, \ldots\}$.
- Parameter space: $\{\lambda \in \mathbf{R} \mid \lambda > 0\}$
- Probability mass function on $k \in \mathcal{Y}$:

$$p(k;\lambda) = \lambda^k e^{-\lambda} / (k!)$$



Beta Family

- Support $\mathcal{Y} = (0, 1)$. [The unit interval.]
- Parameter space: $\{\theta = (\alpha, \beta) \mid \alpha, \beta > 0\}$
- Probability density function on $y \in \mathcal{Y}$:

$$p(y; a, b) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}.$$

Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons`http://taps-graph-review.wikispaces.com/Box+and+Whisker+Plots`.

Gamma Family

- Support $\mathcal{Y} = (0, \infty)$. [Positive real numbers]
- Parameter space: $\{\theta = (k, \theta) \mid k > 0, \theta > 0\}$
- Probability density function on $y \in \mathcal{Y}$:

$$p(y; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}.$$

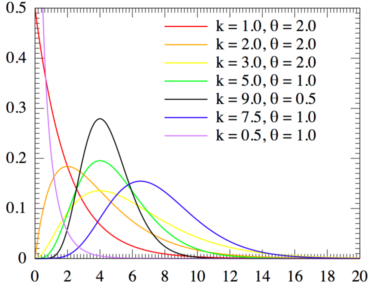


Figure from Wikipedia.

Maximum Likelihood Estimation

Suppose we have a parametric model $\{p(y;\theta) \mid \theta \in \Theta\}$ and a sample $\mathcal{D} = \{y_1, \ldots, y_n\}$.

Definition

The maximum likelihood estimator (MLE) for θ in the model $\{p(y,\theta) \mid \theta \in \Theta\}$ is

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, L_{\mathcal{D}}(\theta) = \underset{\theta \in \Theta}{\arg\max} \prod_{i=1}^{n} p(y_i;\theta).$$

In practice, we prefer to work with the **log likelihood**. Same maximum but

$$\log p(\mathcal{D};\theta) = \sum_{i=1}^{n} \log p(y_i;\theta),$$

and sums are easier to work with than products.

Maximum Likelihood Estimation

- Finding the MLE is an optimization problem.
- For some model families, calculus gives closed form for MLE.
- Can also use numerical methods we know (e.g. SGD).
- Note: In certain situations, the MLE may not exist.
  - But there is usually a good reason for this.
- e.g. Gaussian family $\left\{\mathcal{N}(\mu, \sigma^2 \mid \mu \in \mathbf{R}, \sigma^2 > 0\right\}$, Single observation y.
  - Take $\mu = y$ and $\sigma^2 \to 0$ drives likelihood to infinity. MLE doesn't exist.

Example: MLE for Poisson

- Suppose we've observed some counts $\mathcal{D} = \{k_1, \ldots, k_n\} \in \{0, 1, 2, 3, \ldots\}$.
- The Poisson log-likelihood for a single count is

$$\begin{aligned} \log\left[p(k;\lambda)\right] &= \log\left[\frac{\lambda^k e^{-\lambda}}{k!}\right] \\ &= k\log\lambda - \lambda - \log\left(k!\right) \end{aligned}$$

- The full log-likelihood is

$$\log p(\mathcal{D}, \lambda) = \sum_{i=1}^{n}\left[k_i \log\lambda - \lambda - \log\left(k_i!\right)\right]$$

Example: MLE for Poisson

- The full log-likelihood is

$$\log p(\mathcal{D},\lambda) = \sum_{i=1}^{n}\left[k_i \log\lambda - \lambda - \log\left(k_i!\right)\right]$$

- First order condition gives

$$\begin{aligned} 0 = \frac{\partial}{\partial\lambda}\left[\log p(\mathcal{D},\lambda)\right] &= \sum_{i=1}^{n}\left[\frac{k_i}{\lambda} - 1\right] \\ \Longrightarrow \lambda &= \frac{1}{n}\sum_{i=1}^{n} k_i \end{aligned}$$

- So MLE $\hat{\lambda}$ is just the mean of the counts.

Test Set Log Likelihood for Penn Station, Mon-Fri 7-8pm

Method	Test Log-Likelihood
Poisson	-392.16
Negative Binomial	-188.67
Histogram (Bin width = 7)	$-\infty$
95% Histogram +.05 NegBin	-203.89

Statistical Learning Formulation

Probability Estimation as Statistical Learning

- Output space $\mathcal{Y}$
- **Action space**
  $\mathcal{A} = \{p(y) \mid p \text{ is a probability density or mass function on } \mathcal{Y}\}$.
- How to encode our objective of "high likelihood" as a loss function?
- Define loss function as the negative log-likelihood of y under $p(\cdot)$:

$$\begin{aligned} \ell: \quad \mathcal{A} \times \mathcal{Y} &\to \mathbf{R} \\ (p, y) &\mapsto -\log p(y) \end{aligned}$$

Probability Estimation as Statistical Learning

- If **true** distribution of y is q, then **risk** of predicted distribution p is

$$R(p) = \mathbb{E}_{y \sim q}\left[-\log p(y)\right].$$

- The empirical risk of p for a sample $\mathcal{D} = \{y_1, \ldots, y_n\} \in \mathcal{Y}$ is

$$\hat{R}(p) = -\sum_{i=1}^{n} \log p(y_i),$$

  which is exactly the **negative log-likelihood of p for the data $\mathcal{D}$.**
- Therefore, MLE is just an empirical risk minimizer.

Estimation Distributions, Overfitting, and Hypothesis Spaces

- Just as in classification and regression, MLE (i.e. ERM) can overfit!
- Example Hypothesis Spaces / Probability Models:
  - $\mathcal{F} = \{$Poisson distributions$\}$.
  - $\mathcal{F} = \{$Negative binomial distributions$\}$.
  - $\mathcal{F} =\{$Histogram with 10 bins$\}$
  - $\mathcal{F} =\{$Histogram with bin for every $y \in \mathcal{Y}\}$ [will likely overfit for continuous data]
  - $\mathcal{F} = \{$Depth 5 decision trees with histogram estimates in leaves$\}$
- How to judge with hypothesis space works the best?
- Choose the model with the **highest likelihood for a test set**.

Generalized Regression

Generalized Regression / Conditional Distribution Estimation

- Given X, predict *probability distribution* $p(y \mid x)$
- How do we represent the probability distribution?
- We'll consider *parametric families* of distributions.
    - distribution represented by parameter vector
- Examples:
    1. Logistic regression (Bernoulli distribution)
    2. Probit regression (Bernoulli distribution)
    3. Poisson regression (Poisson distribution)
    4. Linear regression (Normal distribution, fixed variance)
    5. Generalized Linear Models (GLM) (encompasses all of the above)
    6. Generalized Additive Models (GAM)
    7. Gradient Boosting Machines (GBM) / AnyBoost [with likelihood loss function]

Generalized Regression as Statistical Learning

- Input space $\mathcal{X}$
- Output space $\mathcal{Y}$
- All pairs (x, y) are independent with distribution $P_{\mathcal{X}\times\mathcal{Y}}$.
- **Action space**
  $\mathcal{A} = \{p(y) \mid p$ is a probability density or mass function on $\mathcal{Y}\}$.
- Hypothesis spaces contain decision functions $f : \mathcal{X} \to \mathcal{A}$.
  - Given an $x \in \mathcal{X}$, predict a probability distribution $p(y)$ on $\mathcal{Y}$.

A Note on Notation

- Hypothesis spaces contain decision functions $f : \mathcal{X} \to \mathcal{A}$.
    - Given an $x \in \mathcal{X}$, predict a probability distribution $p(y)$ on $\mathcal{Y}$.
- Let f be a decision function.
    - In regression, $f(x) \in \mathbf{R}$
    - In hard classification, $f(x) \in \{-1, 1\}$
    - For generalized regression, $f(x) \in ?$
- $f(x)$ is a PDF or PMF on $\mathcal{Y}$.
- If $p = f(x)$, can evaluate $p(y)$ for predicted probability of y.
- Or just write $[f(x)](y)$ or even $f(x)(y)$.

Generalized Regression as Statistical Learning

- The risk of decision function $f : \mathcal{X} \to \mathcal{A}$

$$R(f) = -\mathbb{E}_{x,y} \log [f(x)](y),$$

where $f(x)$ is a PDF or PMF on $\mathcal{Y}$, and we're evaluating it on Y.

- The empirical risk of f for a sample $\mathcal{D} = \{y_1, \ldots, y_n\} \in \mathcal{Y}$ is

$$\hat{R}(f) = -\sum_{i=1}^{n} \log [f(x_i)](y_i).$$

This is called the negative **conditional log-likelihood**.

Bernoulli Regression

Probabilistic Binary Classifiers

- Setting: $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{0,1\}$
- For each x, need to predict a distribution on $\mathcal{Y} = \{0,1\}$.
- What kind of parametric distribution could be supported on $\{0,1\}$?
- Not a lot of choices....
- Bernoulli!
- For each x,
  - predict the Bernoulli parameter $\theta = p(y = 1 \mid x)$.

Linear Probabilistic Classifiers

- Setting: $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{0, 1\}$
- Want prediction function $x \mapsto \theta = p(y = 1 \mid x)$.
- We need $\theta \in [0, 1]$.
- For a "linear method", we can write this in two steps:

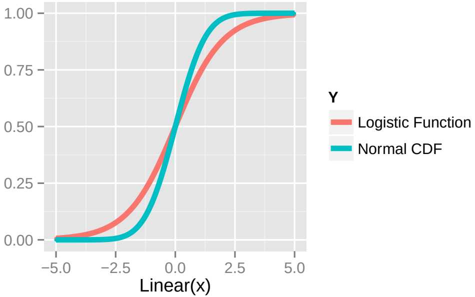
$$\underbrace{x}_{\in \mathbf{R}^D} \mapsto \underbrace{w^T x}_{\in \mathbf{R}} \mapsto \underbrace{f(w^T x)}_{\in [0,1]},$$

  where $f : \mathbf{R} \to [0, 1]$ is called the **transfer** or **inverse link** function.
- Probability model is then

$$p(y = 1 \mid x) = f(w^T x)$$

Inverse Link Functions

- Two commonly used "inverse link" functions to map from $w^T x$ to θ:



- Logistic function $\Longrightarrow$ Logistic Regression
- Normal CDF $\Longrightarrow$ Probit Regression

Learning

- $\mathcal{X} = \mathbf{R}^d$
- $\mathcal{Y} = \{0, 1\}$
- $\mathcal{A} =, 1$ (Representing Bernoulli(θ) distributions by $\theta \in [0, 1]$)
- $\mathcal{H} = \left\{x \mapsto f\left(w^T x\right) \mid w \in \mathbf{R}^d\right\}$
- We can choose w using maximum likelihood...

Bernoulli Regression: Likelihood Scoring

- Suppose we have data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Compute the model likelihood for $\mathcal{D}$:

$$\begin{aligned} p_w(\mathcal{D}) &= \prod_{i=1}^{n} p_w(y_i \mid x_i) \text{ [by independence]} \\ &= \prod_{i=1}^{n} \left[f(w^T x_i)\right]^{y_i} \left[1 - f(w^T x_i)\right]^{1-y_i}. \end{aligned}$$

- Huh? Remember $y_i \in \{0, 1\}$.
- Easier to work with the log-likelihood:

$$\log p_w(\mathcal{D}) = \sum_{i=1}^{n} y_i \log f(w^T x_i) + (1 - y_i) \log \left[1 - f(w^T x_i)\right]$$

Bernoulli Regression: MLE

- Maximum Likelihood Estimation (MLE) finds w maximizing $\log p_w(\mathcal{D})$.
- Equivalently, minimize the objective function

$$J(w) = -\left[\sum_{i=1}^{n} y_i \log f(w^T x_i) + (1 - y_i) \log\left[1 - f(w^T x_i)\right]\right]$$

- For differentiable f,
  - $J(w)$ is differentiable, and we can use our standard tools.
- Homework: Derive the SGD step directions for logistic regression.

Multinomial Logistic Regression

Multinomial Logistic Regression

- Setting: $\mathcal{X} = \mathbf{R}^d$, $\mathcal{Y} = \{1, \ldots, k\}$
- The numbers $(\theta_1, \ldots, \theta_k)$ where $\sum_{c=1}^{k} \theta_c = 1$ represent a
  - "**multinoulli**" or "**categorical**" distribution.
- For each x, we want to produce a distribution on the k classes.
- That is, for each x and each $y \in \{1, \ldots, y\}$, we want to produce a probability

$$p(y \mid x) = \theta_y,$$

where $\sum_{y=1}^{K} \theta_y = 1$.

Multinomial Logistic Regression: Classic Setup

- From each x, we compute a linear score function for each class:

$$x \mapsto (\langle w_1, x\rangle, \ldots, \langle w_k, x\rangle) \in \mathbf{R}^k$$

- We need to map this $\mathbf{R}^k$ vector into a probability vector.
- Use the **softmax function:**

$$(\langle w_1, x\rangle, \ldots, \langle w_k, x\rangle) \mapsto \left(\frac{\exp\left(w_1^T x\right)}{\sum_{c=1}^{K} \exp\left(w_c^T x\right)}, \ldots, \frac{\exp\left(w_k^T x\right)}{\sum_{c=1}^{K} \exp\left(w_c^T x\right)} \right)$$

- If $\theta \in \mathbf{R}^k$ is the output of the softmax, note that

$$\begin{aligned} \theta_i &> 0 \\ \sum_{i=1}^{k} \theta_i &= 1 \end{aligned}$$

Multinomial Logistic Regression: Classic Setup

- Putting this together, we write multinomial logistic regression as

$$p(y \mid x) = \frac{\exp\left(w_y^T x\right)}{\sum_{c=1}^{K} \exp\left(w_c^T x\right)},$$

  where we've introduced parameter vectors $w_1, \ldots, w_k \in \mathbf{R}^d$.
- Do we still see score functions in here?
- Can view $x \mapsto w_y^T x$ as the score for class y, for $y \in \{1, \ldots, k\}$.
- We can also "flatten" this as we did for multiclass classification.
  - Introduce a class-sensitive feature vector $\Psi(x, y) \in \mathbf{R}^{d \times k}$
  - Parameter vector $w \in \mathbf{R}^{d \times k}$.
-

Poisson Regression

Poisson Regression: Setup

- Input space $\mathcal{X} = \mathbf{R}^d$, Output space $\mathcal{Y} = \{0, 1, 2, 3, 4, \ldots\}$
- Hypothesis space consists of functions $f : x \mapsto \text{Poisson}\left(\lambda(x)\right)$.
  - That is, for each x, $f(x)$ returns a Poisson with mean $\lambda(x) \in (0, \infty)$.
  - What function?
- Recall $\lambda > 0$.
- In Poisson regression, x enters **linearly:** $x \mapsto w^T x \mapsto \lambda = f(w^T x)$.
- Standard approach is to take

$$\lambda(x) = \exp\left(w^T x\right),$$

  for some parameter vector w.
- Note that range of $\lambda(x) = (0, \infty)$, (appropriate for the Poisson parameter).

Poisson Regression: Likelihood Scoring

- Suppose we have data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Recall the log-likelihood for Poisson is:

$$\log p(\mathcal{D}, \lambda) = \sum_{i=1}^{n} [y_i \log \lambda - \lambda - \log(y_i!)]$$

- Plugging in $\lambda(x) = \exp\left(w^T x\right)$, we get

$$\begin{aligned}\log p(\mathcal{D}, \lambda) &= \sum_{i=1}^{n} \left[y_i \log\left[\exp\left(w^T x\right)\right] - \exp\left(w^T x\right) - \log(y_i!)\right] \\ &= \sum_{i=1}^{n} \left[y_i w^T x - \exp\left(w^T x\right) - \log(y_i!)\right]\end{aligned}$$

- Maximize this w.r.t. w to find the Poisson regression.
- No closed form for optimum, but it's concave, so easy to optimize.

Conditional Gaussian Regression

Gaussian Regression

- Input space $\mathcal{X} = \mathbf{R}^d$, Output space $\mathcal{Y} = \mathbf{R}$
  - Hypothesis space consists of functions $f : x \mapsto \mathcal{N}\left(w^T x, \sigma^2\right)$.
  - For each x, $f(x)$ returns a particular Gaussian density with variance σ^2 .
  - Choice of w determines the function.
- For some parameter $w \in \mathbf{R}^d$, can write our prediction function as

$$[f_w(x)](y) = p_w(y \mid x) = \mathcal{N}(y \mid w^T x, \sigma^2),$$

  where $\sigma^2 > 0$.
- Given some i.i.d. data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, how to assess the fit?

Gaussian Regression: Likelihood Scoring

- Suppose we have data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$.
- Compute the model likelihood for $\mathcal{D}$:

$$p_w(\mathcal{D}) = \prod_{i=1}^{n} p_w(y_i \mid x_i) \text{ [by independence]}$$

- Maximum Likelihood Estimation (MLE) finds w maximizing $p_w(\mathcal{D})$.
- Equivalently, maximize the data log-likelihood:

$$w^* = \underset{w \in \mathbf{R}^d}{\arg\max} \sum_{i=1}^{n} \log p_w(y_i \mid x_i)$$

- Let's start solving this!

Gaussian Regression: MLE

- The conditional log-likelihood is:

$$
\begin{aligned}
& \sum_{i=1}^{n} \log p_w(y_i \mid x_i) \\
= & \sum_{i=1}^{n} \log \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right) \right] \\
= & \underbrace{\sum_{i=1}^{n} \log \left[\frac{1}{\sigma\sqrt{2\pi}} \right]}_{\text{independent of } w} + \sum_{i=1}^{n} \left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2} \right)
\end{aligned}
$$

- MLE is the w where this is maximized.
- Note that σ^2 is irrelevant to finding the maximizing w.
- Can drop the negative sign and make it a minimization problem.

Gaussian Regression: MLE

- The MLE is

$$w^* = \underset{w \in \mathbf{R}^d}{\arg\min} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

- This is exactly the objective function for least squares.
- From here, can use usual approaches to solve for w^*(linear algebra, calculus, iterative methods etc.)
- NOTE: Parameter vector w only interacts with x by an inner product

Generalized Linear Models (Lite)

Natural Exponential Families

- $\{p_\theta(y) \mid \theta \in \Theta \subset \mathbf{R}^d\}$ is a family of pdf's or pmf's on $\mathcal{Y}$.
- The family is a **natural exponential family** with parameter θ if

$$p_\theta(y) = \frac{1}{Z(\theta)} h(y) \exp\left[\theta^T y\right].$$

- $h(y)$ is a **nonnegative** function called the **base measure.**
- $Z(\theta) = \int_{\mathcal{Y}} h(y) \exp\left[\theta^T y\right]$ is the **partition function**.
- The **natural parameter space** is the set $\Theta = \{\theta \mid Z(\theta) < \infty\}$.
    - the set of θ for which $\exp\left[\theta^T y\right]$ can be normalized to have integral 1
- θ is called the **natural parameter**.
- Note: In exponential family form, family typically has a different parameterization than the "standard" form.

Specifying a Natural Exponential Family

- The family is a **natural exponential family** with parameter θ if

$$p_\theta(y) = \frac{1}{Z(\theta)} h(y) \exp\left[\theta^T y\right].$$

- To specify a natural exponential family, we need to choose $h(y)$.
  - Everything else is determined.
- Implicit in choosing $h(y)$ is the choice of the support of the distribution.

Natural Exponential Families: Examples

The following are univariate natural exponential families:

1. Normal distribution with known variance.
2. Poisson distribution
3. Gamma distribution (with known k parameter)
4. Bernoulli distribution (and Binomial with known number of trials)

Example: Poisson Distribution

- For Poisson, we found the log probability mass function is:

$$\log\left[p(y;\lambda)\right] = y\log\lambda - \lambda - \log\left(y!\right).$$

- Exponentiating this, we get

$$p(y;\lambda) = \exp\left(y\log\lambda - \lambda - \log\left(y!\right)\right).$$

- If we reparameterize, taking $\theta = \log\lambda$, we can write this as

$$\begin{aligned} p(y,\theta) &= \exp\left(y\theta - e^{\theta} - \log\left(y!\right)\right) \\ &= \frac{1}{y!}\frac{1}{e^{e^{\theta}}}\exp\left(y\theta\right), \end{aligned}$$

  which is in natural exponential family form, where

$$\begin{aligned} Z(\theta) &= \exp\left(e^{\theta}\right) \\ h(y) &= \frac{1}{y!}. \end{aligned}$$

- $\theta = \log\lambda$ is the **natural parameter**.

Generalized Linear Models [with Canonical Link]

- In GLMs, we first choose a natural exponential family.
  - (This amounts to choosing $h(y)$.)
- The idea is to plug in $w^T x$ for the natural parameter.
- This gives models of the following form:

$$p_\theta(y \mid x) = \frac{1}{Z(w^T x)} h(y) \exp\left[(w^T x) y\right].$$

- This is the form we had for Poisson regression.
- **Note**: This is very convenient, but **only works if $\Theta = \mathbf{R}$.**

Generalized Linear Models [with General Link]

- More generally, choose a function $\psi : \mathbf{R} \to \Theta$ so that

$$x \mapsto w^T x \mapsto \psi(w^T x),$$

where $\theta = \psi(w^T x)$ is the natural parameter for the family.

- So our final prediction (for one-parameter families) is:

$$p_\theta(y \mid x) = \frac{1}{Z(\psi(w^T x))} h(y) \exp\left[\psi(w^T x) y\right].$$