Lecture 5: Overview

Statistics uses exponential families to construct losses and priors

$$P_G(\mathbf{x}|\boldsymbol{\theta}) = e^{\boldsymbol{\theta}\cdot\mathbf{x} - G(\boldsymbol{\theta})} P_0(\mathbf{x})$$
$$-\log P_G(\mathbf{x}|\boldsymbol{\theta}) = G(\boldsymbol{\theta}) - \boldsymbol{\theta}\cdot\mathbf{x} + \text{const}$$

Convex because $G(\boldsymbol{\theta})$ convex

Machine Learning uses Bregman divergences to construct losses & regularizers

- More general than exponential families
- Convex by design
- Relative entropy between two distribution of the same family w. different parameters is Bregman divergence (later)

Training parameters

- Online - get one example at a time or examples do not fit
  - training based on loss of single example
  - need inertia term
- Batch - training based on loss of whole batch
- Mini batch

Losses often incorporate a regularization term

- First order updates: use only gradient info
- Second order update: use gradient and Hessian info

  Closed form solutions for linear regression

Stochastic gradient descent

- Online update based on random example

  (mini-batch of random examples)