Conjugate Priors
Bayesian Linear Regression
Type II Maximum Likelihood

- Project Proposal due now ✓
  [CPSC 540]
- A6 due Friday (tutorials or my mailbox in 201)
  * please take an assignment to grade.
- Midterm Monday
  * study guide on Piazza
  * you are allowed 2 pages of notes
  * tutorials this week
  * Auditors/crashes:
    X836 3p: Dimitri Bertsekas

Learning Principles

	Train
Maximum likelihood	$\hat{h} = \arg\max_h p(D \mid h)$
MAP, posterior mode	$\hat{h} = \arg\max_h p(h \mid D, \lambda) \propto p(D \mid h)\, p(h \mid \lambda)$
Bayesian	$p(h \mid D, \lambda) = \dfrac{p(D \mid h)\, p(h \mid \lambda)}{p(D \mid \lambda)}$

Last time:

beta $\propto$ (Bernoulli) (beta)

(posterior) $\propto$ (likelihood) (prior)

Predict

$p(\hat{D} \mid \hat{h})$

$p(\hat{D} \mid \hat{h})$

$$p(\hat{D} \mid D, \lambda) = \int p(\hat{D} \mid h)\, p(h \mid D, \lambda)\, dh$$

or $p(\hat{D} \mid h_{mean}) \longrightarrow$ minimizes L_2 risk

or $p(\hat{D} \mid h_{median}) \longrightarrow$ minimizes absolute risk.

Conjugate Priors

Beta prior is "conjugate" to Bernoulli likelihood

— with these choices, prior and posterior are fro

— makes the integrals "easy"

Likelihood		Conjugate prior
Bernoulli, binomial		beta
multinoulli, multinomial		Dirichlet
Gaussian	$p(x\|\mu)$	Gaussian
Gaussian	$p(x\|\sigma^2)$	inverse gamma
	$p(x\|\mu,\sigma^2)$	normal inverse gamma
Gaussian	$p(x\|\Sigma)$	inverse Wishart
Gaussian	$p(x\|\mu,\Sigma)$	normal inverse Wishart

(inverse Ch
is sp

posterior predictive is student 't'

(heavier tailed than Gaussian/Laplace)

O

G

G

G

$$\int\int N(x|\mu,\Sigma)\, NIW(\mu,\Sigma|m,\kappa,\nu,S)\, d\mu\, d\Sigma$$

family.

(*) If you don't have conj. priors:

- mixture of conjugate priors
- numerical integration (low bins)
- variational methods
  (Laplace, mean-field, variational Bayes,
  loopy belief propagation,
  expectation propagation)
- Monte Carlo
  (Gibbs sampling, Metropolis-Hastings,
  Sequential Monte Carlo)

Project
Proposals
↓

Conjugate Priors

Bayesian Linear Regression

Type II Maximum Likelihood

When do conjugate priors exist?

Exponential Family

θ: parameter

x: data

$$p(x \mid \theta) = \frac{1}{Z(\theta)} h(x) \exp(\theta^T \phi(x))$$

Scaling constant (usually 1)

"partition function"

"sufficient statistics"

$$\propto \exp(\theta^T \phi(x) + \log[h(x)])$$

$$A(\theta) = \log Z(\theta)$$

"cumulant function"

$$\frac{dA}{d\theta} = E[\phi(x)]$$

$$\frac{d^2A}{d\theta^2} = \mathrm{var}[\phi(x)]$$

Learning Principles

Train

Maximum likelihood $\hat{h} = \underset{h}{\arg\max}\ p(D|h)$ ✓

MAP, posterior mode $\hat{h} = \underset{h}{\arg\max}\ p(h|D, \underline{\lambda}) \propto p(D|h)p($

Bayesian $p(h|D, \lambda) = \dfrac{p(D|h)p(h|\lambda)}{p(D|\lambda)}$

$$NLL(\theta) = -\theta^T \phi(x) + \underbrace{\log(Z(\theta))}_{\text{log-sum-exp}} + \text{const.}$$

$$\nabla NLL(\theta) = -\phi(x) + E[\phi(x)]$$

At MLE $\phi(x) = E[\phi(x)]$

Bayesian linear regression

$$f(x_i) = w^T x_i$$

$$y = f(x_i) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$$y_i \sim N(f(x_i), \sigma^2)$$

$$y_i \sim N(w^T x_i, \sigma^2) \qquad w_j \sim N(0, \lambda^{-1})$$

$$p(w \mid X, y) \propto \exp\left(\frac{1}{2\sigma^2}(y - Xw)^T(y - Xw)\right)\exp($$

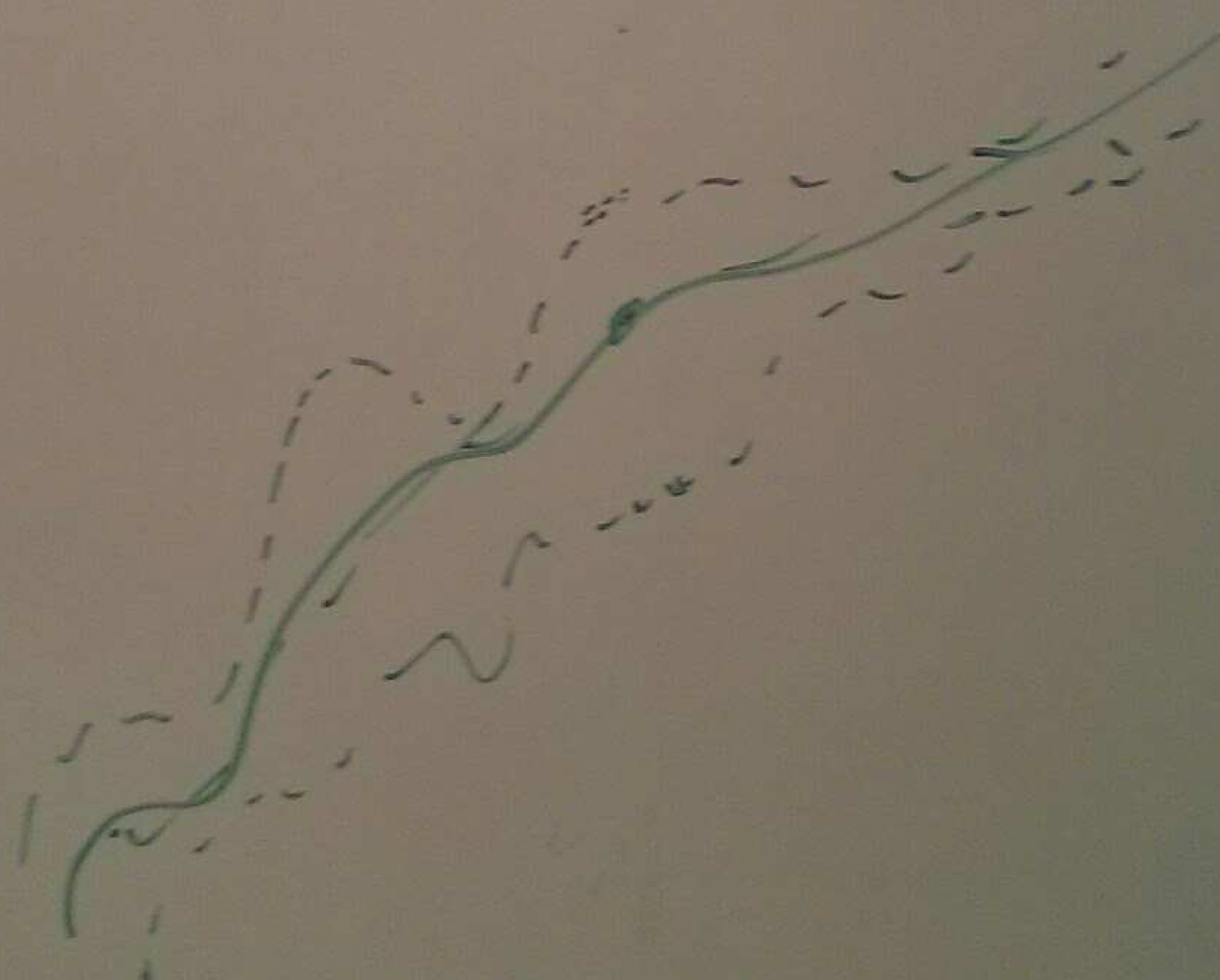
$$\propto \exp\left(-\frac{1}{2}(w - \bar{w})^T \underbrace{\left(\frac{1}{\sigma^2}X^T X + \lambda I\right)}_{\text{"A"}}^{-1} (\right.$$

$$\bar{w} = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}X^T X + \lambda I\right)^{-1} X^T y$$

$$w \mid X, y \sim N\left(\frac{1}{\sigma^2}A^{-1}X^T y, A^{-1}\right)$$

$\hat{f}(\hat{x}, X, Y$

$|^2)$



MAP, posterior mode $\hat{h} = \arg\max_h p(h|D, \underline{\lambda}) \propto p(D|h)p(h|\underline{\lambda})$

Bayesian $p(h|D,\lambda) = \frac{p(D|h)p(h|\lambda)}{p(D|\lambda)}$

Type II Maximum Likelihood

"Empirical Bayes"

$\hat{\lambda} = \arg\max_\lambda \underbrace{p(D|\lambda)}_{\text{marginal likelihood "evidence"}} = \int p(D|h)p(h|\lambda)\,dh$

Type II MAP

"hyper-prior"

"hierachical Bayesian model"

$\hat{\lambda} = \arg\max_\lambda p(D|\lambda)p(\lambda)$

"Very Bayesian"

$p(h,\lambda|D) \propto p(D|h)p(h|\lambda)p(\lambda)$

$p(\dots$

$p(\hat{D}\dots$

$p(D\dots$

or $p(D\dots$

or $p(D\dots$

$p(\hat{D} | h) p(h | D, \lambda) dh$

$\rightarrow$ minimizes L_2 risk

$\rightarrow$ minimizes absolute risk.

Automatic relevance determination

"sparse Bayesian learning"

$y_i \sim N(w^T x_i, \sigma^2)$ $\quad w_j \sim N(0, \lambda_j^{-1})$

$\lambda_j \rightarrow \infty$

for irrelevant features

Relevance Vector machine

- RBF basis
- ARD

ugate Priors

sian Linear Regression

II Maximum Likelihood

Stochastic Processes: Infinite set of random variables $\{x_i\}$

Gaussian Processes

- Stochastic process where all finite combinations have a Gaussian distribution
- Completely specified by

mean function: $m(x) = E[f(x)]$

covariance function: $k(x, x') = E[(m(x) - f(x))(m(x) - f(x'))]$

$f(x) \sim GP(m(x), k(x, x'))$

Example:

$f(x) = \phi(x)^T w, \quad w \sim N(0, \Sigma)$

$$E_w[f(x)] = \phi(x)^T \underbrace{E[w]}_{0} = 0$$

$$E_w[f(x) f(x)^T] = \phi(x)^T E[ww^T] \phi(x')$$
$$= \phi(x)^T \Sigma \phi(x')$$

⊛ Non-parametric Bayes

- Dirichlet process (stick-breaking, CRP))
- Indian buffet process
- Polya-urn process