

# MN Regularization using Parameter Priors

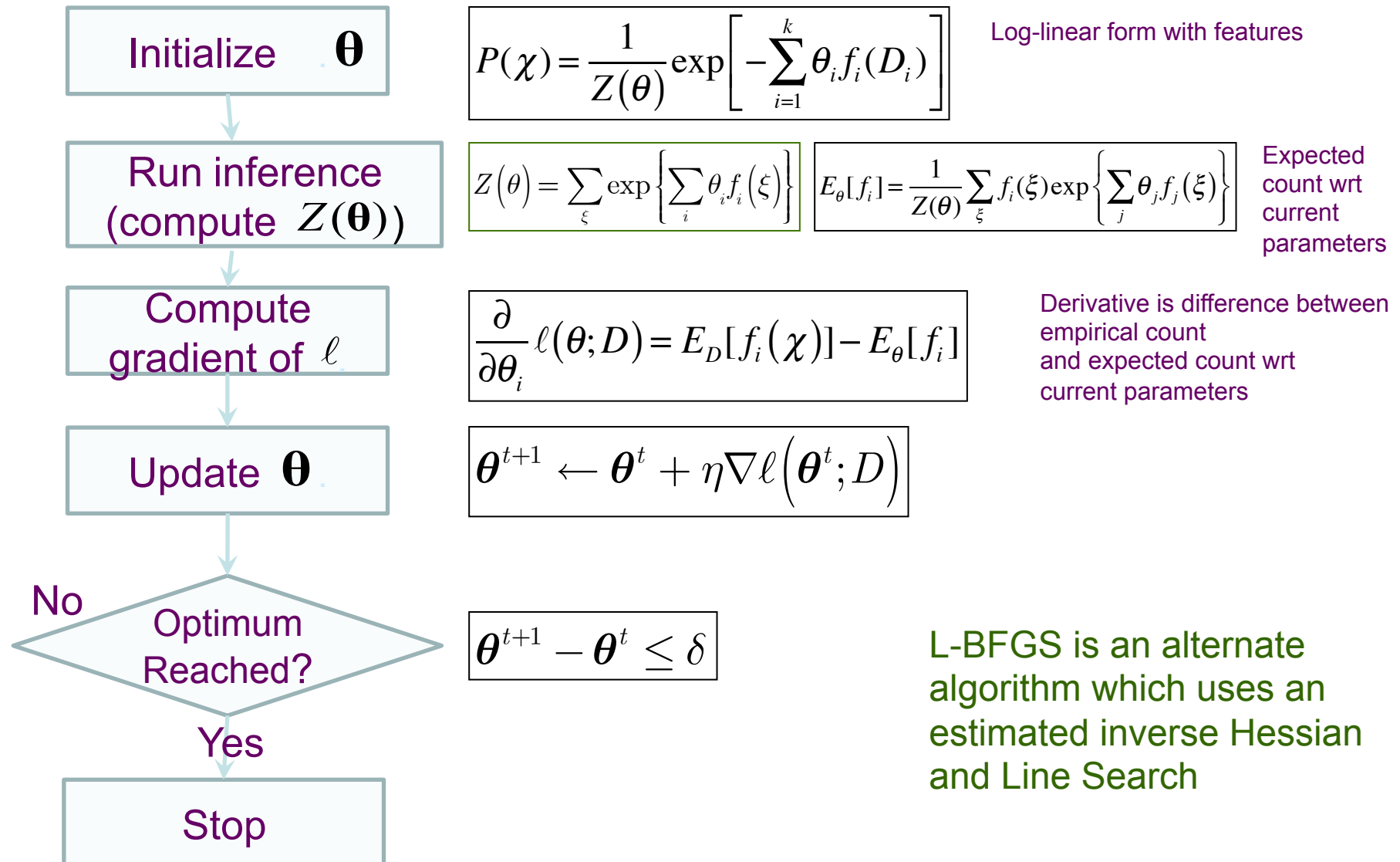
Sargur Srihari

[srihari@cedar.buffalo.edu](mailto:srihari@cedar.buffalo.edu)

# Topics

- Parameter Priors and Regularization
  1. MN parameter estimation methods using ML
  2. Overfitting problem of ML
  3. Local Priors
    1. Gaussian prior and  $L_2$ -regularization
    2. Laplacian prior and  $L_1$ -regularization
  4. Why prefer low-magnitude parameters?
  5. Global Priors

# Iterative ML method for MN params



# Overfitting problem of ML

- Overfitting:
  - Model fits training set exactly; fails to generalize
  - Solution is regularization
- ML estimation is prone to over-fitting
  - True of ML of parameters of MNs as well
    - Not as apparent due to lack of direct correspondence between empirical counts and parameters
  - Overfitting is as much of a problem with MNs
- We can reduce effect of overfitting by:
  - Using a prior distribution over parameters
  - Early stopping, penalize large values, dropout, etc.

# Parameter Priors

- Introduce prior distribution  $P(\boldsymbol{\theta})$ 
  - Defined over model parameters  $\boldsymbol{\theta}$
- Due to non-decomposable likelihood function, a fully Bayesian approach is infeasible

$$P(X_1, \dots, X_n; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp \left\{ \sum_{i=1}^k \theta_i f_i(D_i) \right\}$$

- In a fully Bayesian approach we integrate out the parameters to get the next prediction
- Instead, we perform MAP estimation
  - Find parameters that maximize  $P(\boldsymbol{\theta})P(\mathcal{D} | \boldsymbol{\theta})$

# MAP estimation of $\theta$

- Instead, we perform MAP estimation
- Find parameters that maximize  $P(\theta)P(\mathcal{D} | \theta)$ 
  - where  $P(\theta)$  is the prior distribution
  - and  $\ln P(\mathcal{D} | \theta)$  is expressed in log-space as

$$\ell(\theta : D) = \sum_{i=1} \theta_i \left( \sum_m f_i(\xi[m]) \right) - M \ln Z(\theta)$$

- which in turn is derived from the joint distribution

$$P(X_1, \dots, X_n; \theta) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{i=1}^k \theta_i f_i(D_i) \right\}$$

# Local Prior $P(\theta)$

- Assuming no constraints on conjugacy of prior and likelihood, two commonly used priors:
  1. Gaussian (zero-mean diagonal with equal variances for each of the weights)
    - $L_2$  regularization
      - Weight penalty is quadratic in  $\theta$  (Euclidean norm)
  2. Laplacian distribution (zero-mean)
    - $L_1$  regularization
      - Weight penalty is linear in  $|\theta|$
- Both priors penalize parameters whose magnitude (positive or negative) is large

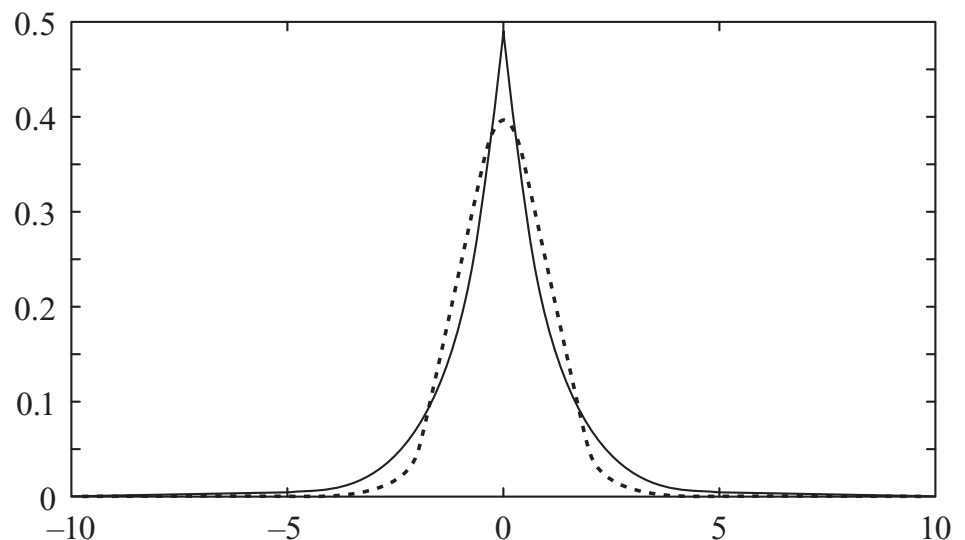
# Gaussian and Laplacian Priors

1. Zero-mean diagonal Gaussian with equal variances for each of the weights

- $L_2$  regularization,
  - penalty is quadratic in  $\theta$

2. Zero-mean Laplacian distribution

- $L_1$  regularization
  - penalty is linear in  $|\theta|$



Gaussian distribution  $\sigma^2=1$   
Laplacian distribution  $\beta=1$



# Gaussian Prior and $L_2$ -Regularization

- Most common is Gaussian prior on log-linear parameters  $\theta$

$$P(\theta | \sigma^2) = \prod_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\theta_i^2}{2\sigma^2}\right\}$$

– For some choice of hyper-parameter (variance)  $\sigma^2$

- Analogous to  $\alpha_i$  in Dirichlet prior for multinomial

- Converting MAP objective  $P(\theta)P(\mathcal{D} | \theta)$  to log-space, gives  $\ln P(\theta) + \ell(\theta : D)$  whose first term

$$\ln P(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^k \theta_i^2$$

– is called an  $L_2$ -regularization term

– Recall  $\frac{1}{M} \ell(\theta : D) = \sum_i \theta_i (E_D[f_i(d_i)]) - \ln Z(\theta)$  and  $\frac{\partial}{\partial \theta_i} \frac{1}{M} \ell(\theta : D) = E_D[f_i(\chi)] - E_\theta[f_i]$

# Laplacian Prior and $L_1$ -Regularization

- Zero-mean Laplacian distribution

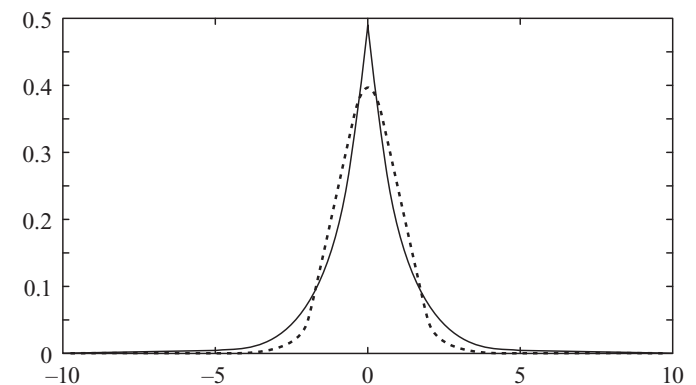
$$P_{Laplacian}(\boldsymbol{\theta}) = \frac{1}{2\beta} \exp\left\{-\frac{|\boldsymbol{\theta}|}{\beta}\right\}$$

- Taking log we obtain a term

$$\ln P(\boldsymbol{\theta}) = -\frac{1}{\beta} \sum_{i=1}^k |\theta_i|$$

– which we wish to minimize

- Generally called  $L_1$ -regularization
- Both forms of regularization penalize parameters whose magnitude is large



Laplacian distribution  $\beta=1$   
Gaussian distribution  $\sigma^2=1$

# Why prefer low magnitude parameters ?

- Properties of prior
  - To pull distribution towards an uninformed one
  - To smooth fluctuations in the data
- A distribution is *smooth* if
  - Probabilities assigned to different assignments are not radically different
- Consider two assignments  $\xi$  and  $\xi'$ 
  - An assignment is an instance of variables  $X_1, \dots, X_n$
- We consider ratio of their probabilities next

# Smoothness resulting from small $\theta$

- Given two assignments  $\xi$  and  $\xi'$ ,
  - Their relative probability is

$$\frac{P(\xi)}{P(\xi')} = \frac{\tilde{P}(\xi) / Z(\theta)}{\tilde{P}(\xi') / Z(\theta)} = \frac{\tilde{P}(\xi)}{\tilde{P}(\xi')}$$

- where the un-normalized probabilities are

$$\tilde{P}(\xi) = \exp \left\{ \sum_{i=1}^k \theta_i f_i(\xi) \right\}$$

- In log-space, log-probability ratio is

$$\ln \frac{P(\xi)}{P(\xi')} = \sum_{i=1}^k \theta_i f_i(\xi) - \sum_{i=1}^k \theta_i f_i(\xi') = \sum_{i=1}^k \theta_i (f_i(\xi) - f_i(\xi'))$$

- When  $\theta_i$ 's have small magnitude, this log-ratio is also bounded, i.e., probabilities are similar
- This results in a smooth distribution

# Comparison of $L_1$ and $L_2$ Regularization

- Both  $L_1$  and  $L_2$  penalize parameter magnitude
  - Encode belief that model weights should be small (Close to zero)
- In Gaussian case ( $L_2$ ) , penalty grows quadratically with parameters
  - An increase in  $\theta_i$  from 3 to 3.1 is penalized more than  $\theta_i$  from 0 to 0.1
  - Leads to many small parameters
- In Laplacian case ( $L_1$ ), penalty grows linearly
  - Results in fewer edges and is more tractable

# Efficiency of Optimization

- Both  $L_1$ - and  $L_2$ - Regularization terms are Concave

— i.e.,  $\boxed{\ln P(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2} \sum_{i=1}^k \theta_i^2}$  and  $\boxed{\ln P(\boldsymbol{\theta}) = -\frac{1}{\beta} \sum_{i=1}^k |\theta_i|}$  are concave

- Because Log-likelihood  $\boxed{\frac{1}{M} \ell(\boldsymbol{\theta} : D) = \sum_i \theta_i (E_D[f_i(d_i)]) - \ln Z(\boldsymbol{\theta})}$  is also Concave, resulting posterior  $\ln P(\boldsymbol{\theta}) + \ell(\boldsymbol{\theta} : D)$  is also concave

- Can be optimized using gradient descent methods
- Introduction of penalty terms eliminates multiple equivalent minima

# Choice of Hyper-parameters

- Regularization hyper-parameters are  $\sigma^2 (L_1)$  and  $\beta (L_2)$ 
  - Larger hyper-parameters mean broader prior
  - Choice of prior has effect on learned model
- Standard method of selecting this parameter is via cross-validation
  - Repeatedly partition training set
    - Learn model over one part with some choice of parameters
    - ensure the performance on the held-out fragment