Approximation and Inference Using Latent Variable Sparse Linear Models

David Wipf, Srikantan Nagarajan
University of California, San Francisco

Jason Palmer, Bhaskar Rao, Kenneth Kreutz-Delgado
University of California, San Diego

NIPS Workshop, 7 December 2007

Overview

- Sparse Linear Models
- Latent variable representations of sparse priors lead to four possibilities:
  1. Standard MAP estimation of unknown model weights
  2. MAP estimation of model hyperparameters.
  3. Local variational approximation (convex lower-bounding)
  4. Global variational approximation (variation Bayes)
- *Unification*: All of these are special cases of hyperparameter MAP estimation using different implicit hyperpriors.
- Choosing the hyperprior, two applications:
  - Finding maximally sparse representations (e.g., sparse coding, RVMs)
  - Active learning, experimental design
- *Extensions*: non-negativity constraints, covariance component estimation, classification.

Sparse Linear Model

- Linear generative model:

$$\mathbf{t} = \Phi\mathbf{w} + \boldsymbol{\varepsilon}$$

Observed n-dimensional data vector

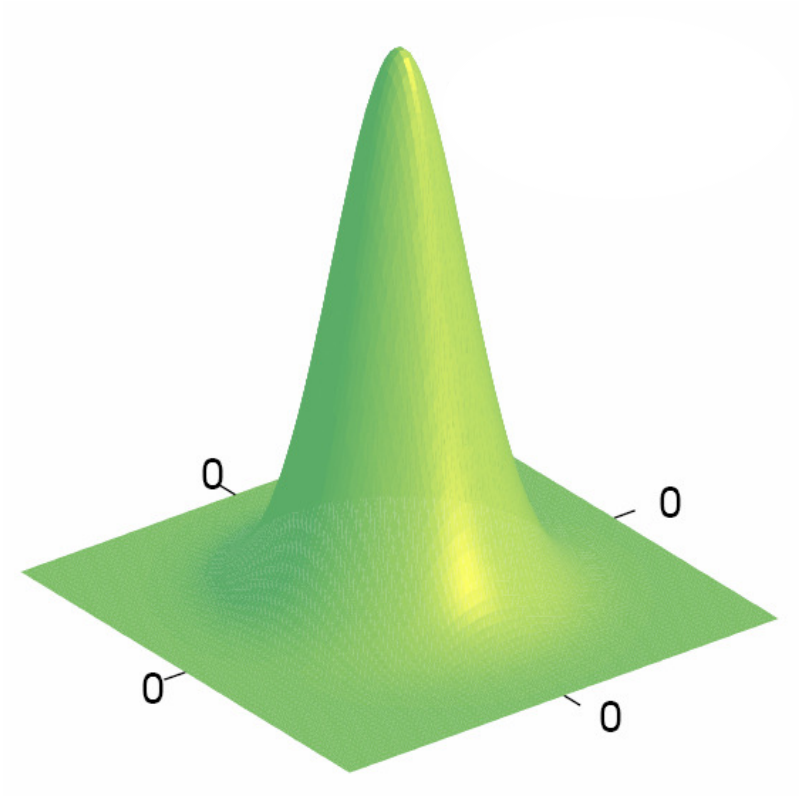
Matrix of m feature vectors

- **Data likelihood**: $p(\mathbf{t} \mid \mathbf{w}; \lambda) \propto \exp\left(-\frac{1}{2\lambda}\left\|\mathbf{t} - \Phi\mathbf{w}\right\|_2^2\right)$

- **Sparse prior**: $p(\mathbf{w}) = \prod_{i=1}^{m} p(w_i) \propto \prod_{i=1}^{m} \exp\left[-g\left(w_i^2\right)\right]$

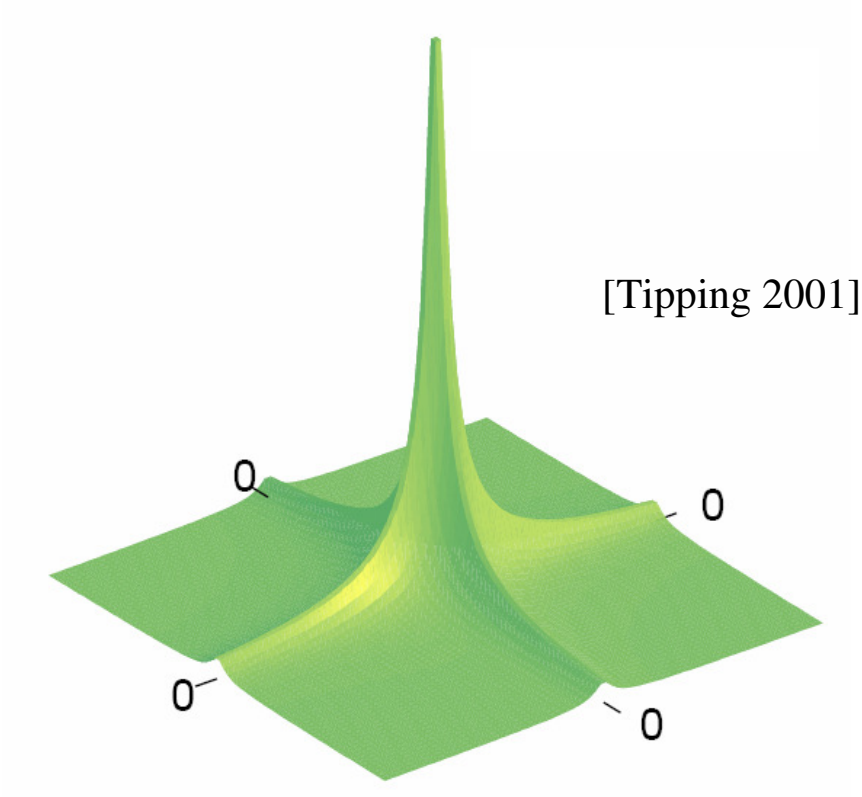
non-decreasing, concave function

Sparse Prior: 2D Example



Gaussian Distribution

[Tipping 2001]



Sparse Distribution

Practical Issues

- Difficulties with joint distribution p(**w**,**t**) = p(**t**|**w**)p(**w**):
  - Often highly multimodal
  - Difficult to evaluate where predominate mass is located.
  - Intractable normalization, cannot compute:

$$p(\mathbf{t}) = \int p(\mathbf{t} \mid \mathbf{w})\, p(\mathbf{w}) d\mathbf{w}$$

- **Conclusion**: Approximations to p(**w**|**t**) and p(**t**) are needed.

Latent Variable Models of Sparse Priors

1. Gaussian scale mixture:

$$p(w_i) = \int N\left(w_i;\ 0, \gamma_i\right) p(\gamma_i) d\gamma_i$$

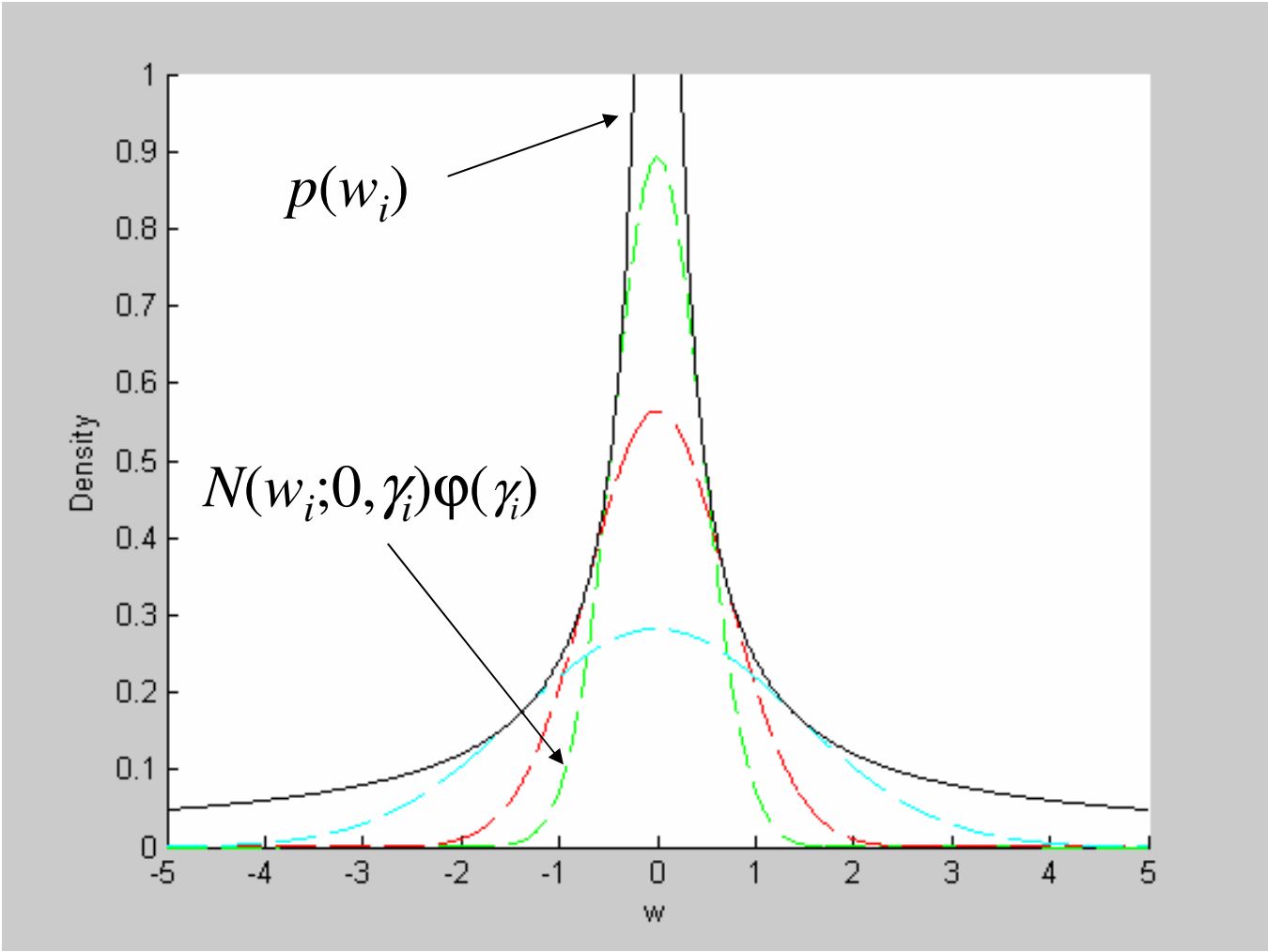
2. Convexity-based representation:

$$p(w_i) = \sup_{\gamma_i \geq 0} N\left(w_i;\ 0, \gamma_i\right) \varphi(\gamma_i)$$

Properties:

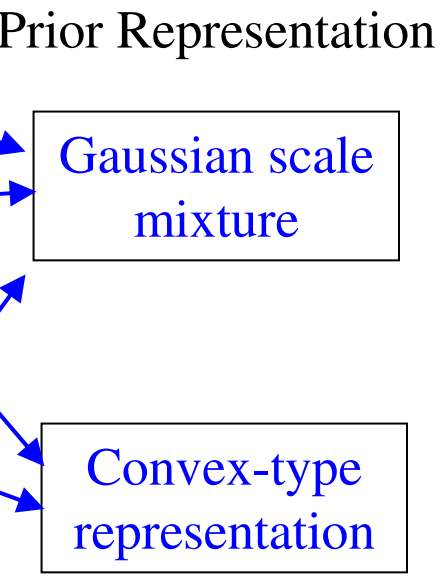
- Essentially all sparse priors can be represented in both forms [Palmer et al., 2006].
- For non-negative functions $p(\gamma_i)$ and $\varphi(\gamma_i)$, resulting $g\left(w_i^2\right)$ will be non-decreasing, concave (sparse).

1d Example of Convex-Type Representation



Four Possibilities for Approximation

1. Standard MAP estimation of $\mathbf{w}$
2. MAP estimation of $\boldsymbol{\gamma}$
3. Local variational approximation, convex bounding
4. Global variational approximation, variational Bayes (VB)



Method I: w-MAP

- **Solve**:

$$\begin{aligned} \mathbf{w}_{\text{MAP}} &= \arg\max_{\mathbf{w}} p(\mathbf{w}\,|\,\mathbf{t}) \\ &= \arg\min_{\mathbf{w}} -\log p(\mathbf{t}\,|\,\mathbf{w})\,p(\mathbf{w}) \\ &= \arg\min_{\mathbf{w}} \tfrac{1}{\lambda}\left\|\mathbf{t}-\Phi\mathbf{w}\right\|_2^2 + \sum_i g\left(w_i^2\right) \end{aligned}$$

- **Approximation**: $p\left(\mathbf{w}\,|\,\mathbf{t}\right) \approx N\left(\mathbf{w};\boldsymbol{\mu}=\mathbf{w}_{\text{MAP}},\Sigma=0\right)$

Method II: γ-MAP

- **Solve**:

$$\boldsymbol{\gamma}_{\text{MAP}} = \arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{\gamma}\,|\,\mathbf{t}), \qquad \boldsymbol{\gamma} = [0,\ldots,\gamma_m]^T$$

$$= \arg\min_{\boldsymbol{\gamma}} -\log \int p(\mathbf{t}\,|\,\mathbf{w}) N(\mathbf{w};\ 0,\boldsymbol{\gamma})\, p(\boldsymbol{\gamma}) d\mathbf{w}$$

$$= \arg\min_{\boldsymbol{\gamma}} \mathbf{t}^T \left(\lambda I + \Phi\Gamma\Phi^T\right)^{-1} \mathbf{t} + \log\left|\lambda I + \Phi\Gamma\Phi^T\right| + \sum_i f(\gamma_i)$$

with $\Gamma = \text{diag}[\boldsymbol{\gamma}], \quad f(\gamma_i) = -\log p(\gamma_i)$

- **Approximation**: $p(\mathbf{w}\,|\,\mathbf{t}) \approx N(\mathbf{w};\boldsymbol{\mu},\Sigma)$

$$\boldsymbol{\mu} = \text{E}[\mathbf{w}\,|\,\mathbf{t},\boldsymbol{\gamma}_{\text{MAP}}] = \lambda^{-1}\Sigma\Phi^T\mathbf{t}$$

$$\Sigma = \text{Cov}[\mathbf{w}\,|\,\mathbf{t},\boldsymbol{\gamma}_{\text{MAP}}] = \left(\lambda^{-1}\Phi^T\Phi + \Gamma_{\text{MAP}}^{-1}\right)^{-1}$$

Method III: Convex Bounding

- The convex-type representation gives the following lower bound:

$$p(w_i) \;\geq\; p(w_i;\gamma_i) \;=\; N\left(w_i;0,\gamma_i\right)\varphi\left(\gamma_i\right)$$

- This gives a family of approximate distributions:

$$p\left(\mathbf{w},\mathbf{t}\right) \approx p\left(\mathbf{w},\mathbf{t};\boldsymbol{\gamma}\right) = p\left(\mathbf{t}\,|\,\mathbf{w}\right)\prod_i p\left(w_i;\gamma_i\right)$$

- **Solve**:

$$\boldsymbol{\gamma}_{\text{OPT}} \;=\; \arg\min_{\boldsymbol{\gamma}} \int \left| p\left(\mathbf{w},\mathbf{t}\right) - p\left(\mathbf{w},\mathbf{t};\boldsymbol{\gamma}\right)\right| d\mathbf{w}$$

$$\equiv\; \arg\max_{\boldsymbol{\gamma}} \int p\left(\mathbf{t}\,|\,\mathbf{w}\right) p\left(\mathbf{w};\boldsymbol{\gamma}\right) d\mathbf{w}$$

- **Approximation**: Same as $\boldsymbol{\gamma}$-MAP

Method IV: Variational Bayes

- The posterior of the complete data $p(\mathbf{w},\boldsymbol{\gamma}|\mathbf{t})$ is typically multimodal and intractable.

- But we can form an approximate, factorial posterior [Attias 2000; Beal 2003]:

$$p(\mathbf{w},\boldsymbol{\gamma}\,|\,\mathbf{t}) \approx \hat{p}(\mathbf{w}\,|\,\mathbf{t})\,\hat{p}(\boldsymbol{\gamma}\,|\,\mathbf{t})$$

- **Solve**:

$$\hat{p}(\mathbf{w}\,|\,\mathbf{t}),\ \hat{p}(\boldsymbol{\gamma}\,|\,\mathbf{t}) \;=\; \arg\min_{q(\mathbf{w}),\,q(\boldsymbol{\gamma})} \mathrm{KL}\left[q(\mathbf{w})q(\boldsymbol{\gamma}) \,\|\, p(\mathbf{w},\boldsymbol{\gamma}\,|\,\mathbf{t})\right]$$

- **Approximation**: $p(\mathbf{w}\,|\,\mathbf{t}) \approx \hat{p}(\mathbf{w}\,|\,\mathbf{t}) = N\left(\mathbf{w};\boldsymbol{\mu},\Sigma\right)$

$\boldsymbol{\mu},\Sigma$ same as before, only with $\boldsymbol{\gamma}_{\mathrm{OPT}} = E_{\hat{p}(\boldsymbol{\gamma}|\mathbf{t})}\left[\boldsymbol{\gamma}\right]$

Unification

- Four different methods, but how different are they?

- If we allow for improper hyperpriors $p(\boldsymbol{\gamma})$, then by inspection $\boldsymbol{\gamma}$-MAP and convex bounding give equivalent approximate distributions when we set $p(\boldsymbol{\gamma}) = \varphi(\boldsymbol{\gamma})$.
  - Note: The associated 'true' posteriors will generally be different.

- VB and convex bounding give equivalent approximations to a fixed $p(\mathbf{w}|\mathbf{t})$ [Palmer et al. 2006].
  - Note: The underlying $p(\boldsymbol{\gamma}) \neq \varphi(\boldsymbol{\gamma})$.

Conclusion:

$\boldsymbol{\gamma}$-MAP encompasses both VB and convex bounding.

Unification Cont.

- What about **w**-MAP?

- **Result** [Wipf et al. 2007]:
  - Let $\mathbf{w}_{\text{MAP}}$ be the solution to an abitrary **w**-MAP problem.
  - Then there exists a limiting $\boldsymbol{\gamma}$-MAP problem that produces the posterior approximation:

$$p\left(\mathbf{w} \mid \mathbf{t}, \boldsymbol{\gamma}_{\text{MAP}}\right) \;=\; N\left(\mathbf{w}; \boldsymbol{\mu} = \mathbf{w}_{\text{MAP}},\; \Sigma = 0\right)$$

- **Note**: The associated 'true' posterior and underlying $p(\boldsymbol{\gamma})$ will both be very different.

Choosing a Model

- **Summary**:
  - 4 different approximation strategies can all be viewed as special cases of hyperparameter MAP.
  - Only real distinction is choice of $f(\boldsymbol{\gamma}) = -\log p(\boldsymbol{\gamma})$.

- **General Considerations for choosing $f(\boldsymbol{\gamma})$**:
  - When possible, better to pick a good $f(\boldsymbol{\gamma})$ than pick a plausible $p(\mathbf{w})$ and work backwards.
  - Ultimate goal is a good posterior approximation; if an absurd 'full' model leads to a very useful approximation, so be it.

- **Specific Issues:**
  1. Optimization
  2. Application-dependent requirements.

Optimization Issues

- If $f(\boldsymbol{\gamma})$ is concave, can implement $\boldsymbol{\gamma}$-MAP using efficient re-weighted L_1 minimization procedure [Wipf and Nagarajan, 2007].

- If $f(\boldsymbol{\gamma}) = -\log p(\boldsymbol{\gamma})$ is not available in closed form, but we can compute

$$g'\left(w_i^2\right) = \frac{\partial g\left(w_i^2\right)}{\partial w_i^2}$$

  then simple EM algorithm exists [Palmer et al. 2006]:

$$\text{E-step:} \quad \text{Compute } \boldsymbol{\mu}, \Sigma$$

$$\text{M-step:} \quad \gamma_i \rightarrow 2g'\left(w_i^2\right)\Big|_{w_i^2 = \mu_i^2 + \Sigma_{ii}}$$

- Greedy methods also exist for some special cases [Tipping and Faul 2003].

Example Applications

1. Finding maximally sparse representations from overcomplete dictionaries of features (e.g., sparse coding, RVMs).

2. Active learning, experimental design (e.g., finding non-random projections for compressed sensing).

Maximally Sparse Representations

- Noiseless case ($\boldsymbol{\varepsilon} = 0$):

$$\mathbf{w}_0 = \arg\min_{\mathbf{w}} \|\mathbf{w}\|_0 \quad \text{s.t.} \quad \mathbf{t} = \Phi\mathbf{w}$$

# of nonzero elements in **w**

- Noisy case ($\boldsymbol{\varepsilon} > 0$):

$$\mathbf{w}_0(\lambda) = \arg\min_{\mathbf{w}} \|\mathbf{t} - \Phi\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_0$$

- **Problem**: Forward model is linear, the inverse problem is very difficult to solve for two reasons:

1. Combinatorial number of local minima
2. Objective is discontinuous

Example

$$\mathbf{t} = \begin{bmatrix} -4 \\ -5 \\ 3 \end{bmatrix}, \qquad \Phi = \begin{bmatrix} 1 & 4 & 1 & 1 & 6 \\ -2 & 1 & -4 & 2 & -3 \\ 3 & 3 & 2 & -2 & 1 \end{bmatrix}$$

Want to find a $\mathbf{w}$ that solves

$$\mathbf{t} = \Phi\,\mathbf{w}$$

non-sparse

$$\mathbf{w} = \begin{bmatrix} 4 \\ -1 \\ 3 \\ 5 \\ -2 \end{bmatrix}$$

sparse

$$\mathbf{w}_0 = \begin{bmatrix} 0 \\ 0 \\ 2 \\ 0 \\ -1 \end{bmatrix}$$

Using γ-MAP to find $\mathbf{w}_0$

1. Choose appropriate $f(\boldsymbol{\gamma})$
2. Compute $\boldsymbol{\gamma}_{\text{MAP}}$
3. Then form $N(\mathbf{w};\boldsymbol{\mu},\Sigma)$ and use $\mathbf{w}_0 \approx \boldsymbol{\mu}$.

Two Criteria for Choosing $f(\gamma)$

Criteria 1:

If w_0 has only one nonzero element, then the associated γ-MAP cost function has a single minimum and $\boldsymbol{\mu} = \mathbf{w}_0$.

In other words, finding only one nonzero element should be very easy.

Two Criteria for Choosing $f(\gamma)$

Criteria 2:

All local minima produce approximations $\boldsymbol{\mu}$ such that

$$\|\boldsymbol{\mu}\|_0 \leq n$$

We should never require more than n nonzero coefficients to represent an n-dimensional signal $\mathbf{t}$.

Result

Over all possible choices for $f(\boldsymbol{\gamma})$, only the selection

$$f(\boldsymbol{\gamma}) = \sum_i \alpha\gamma_i, \qquad \alpha \geq 0$$

satisfies these two criteria [Wipf et al. 2007].

- Sparse Bayesian learning (SBL) [Tipping 2001] and Lasso [Tibshirani 1996] both emerge as special cases:

SBL:	$\alpha \to 0$
Lasso:	$\alpha \to \infty$

Associated ‘Full’ Model

- The associated full distribution will depend on which method we are using:

Implicit Full Model

- *Hyperparameter MAP* } $p(\mathbf{w})$ is Laplacian; $p(\mathbf{w}|\mathbf{t})$ is log-concave, unimodal.

- *Convex lower bounding*
- *Variational Bayes* } $p(\mathbf{w})$ is highly sparse; $p(\mathbf{w}|\mathbf{t})$ is complex, multi-modal

Notes about SBL (and RVMs)

- **Note 1**: Full model assumes

$$p(\gamma_i) \propto \frac{1}{\gamma_i} \quad \Rightarrow \quad p(w_i) \propto \frac{1}{w_i}$$

  Exact inference with this model would be very undesirable.

- **Note 2**: $\boldsymbol{\gamma}$-MAP applied to full model, which implies $f(\gamma_i) = \log \gamma_i$, leads to the trivial solution:

$$\boldsymbol{\gamma}_{\text{MAP}} = 0, \quad \boldsymbol{\mu} = 0, \quad \Sigma = 0,$$

  a good approximation to the full model but not very useful.

- **Note 3**: SBL does $\boldsymbol{\gamma}$-MAP in log space which leads to $f(\boldsymbol{\gamma}) = 0$.

Experimental Design

- **Basic Idea** [Shihao Ji et al. 2007]: Use the approximate posterior

$$p\left(\mathbf{w} \mid \mathbf{t}; \boldsymbol{\gamma}_{\text{MAP}}\right) = N\left(\mathbf{w}; \boldsymbol{\mu}, \Sigma\right)$$

  to learn new rows of the design matrix Φ such that our uncertainty about $\mathbf{w}$ is reduced.

- Choose each additional row to minimize the differential entropy H:

$$H = \tfrac{1}{2}\log\left|\Sigma\right|, \quad \Sigma = \left(\lambda^{-1}\Phi^{T}\Phi + \Gamma_{\text{MAP}}\right)^{-1}$$

Problem

- If $f(\gamma)$ is concave, then at least $m - n$ hyperparameters will be set to exactly zero, regardless of the data.

- When γ_{MAP} is sparse:
  - The posterior mean μ is forced to be sparse, even if the desired posterior mean is far from sparse.
  - There is no posterior uncertainty in the associated zero-valued coefficients; even with limited data the model assigns zero posterior variance.

One Heuristic Solution

- Use $f(\boldsymbol{\gamma})$ from variational relevance vector machine [Bishop and Tipping 2000]:

$$f(\gamma_i) = a \log \gamma_i + \frac{b}{\gamma_i}$$

This prevents any γ_i from going near zero with limited data.

Extensions

1. Non-Negative Sparse Coding
2. Classification
3. Covariance Component Estimation

Non-Negative Sparse Coding

- **Model**:

$$\mathbf{w}_0(\lambda) \quad = \quad \arg\min_{\mathbf{w}\geq 0} \quad \left\|\mathbf{t}-\Phi\mathbf{w}\right\|_2^2 + \lambda\left\|\mathbf{w}\right\|_0$$

- **Problem**: It is intractable to compute:

1. the $\boldsymbol{\gamma}$-MAP cost function

$$p\left(\boldsymbol{\gamma}\,|\,\mathbf{t}\right) \quad = \int p\left(\mathbf{t}\,|\,\mathbf{w};\boldsymbol{\gamma}\right)p(\mathbf{w};\boldsymbol{\gamma})\,p(\boldsymbol{\gamma})d\mathbf{w}$$

2. and the posterior moments $\boldsymbol{\mu}$ and Σ given some $\boldsymbol{\gamma}$.

Non-Negative Sparse Coding Cont.

- It can be shown [Wipf and Nagarajan 2007] that the posterior mean from standard $\boldsymbol{\gamma}$-MAP satisfies

$$\boldsymbol{\mu} \quad = \quad \arg\min_{\mathbf{w}} \tfrac{1}{\lambda}\left\|\mathbf{t}-\Phi\mathbf{w}\right\|_2^2 + g\left(\mathbf{w}\right)$$

where

$$g\left(\mathbf{w}\right) \quad = \quad h^*\left(\left[w_1^2 \cdots w_m^2\right]^T\right)$$

Concave conjugate of

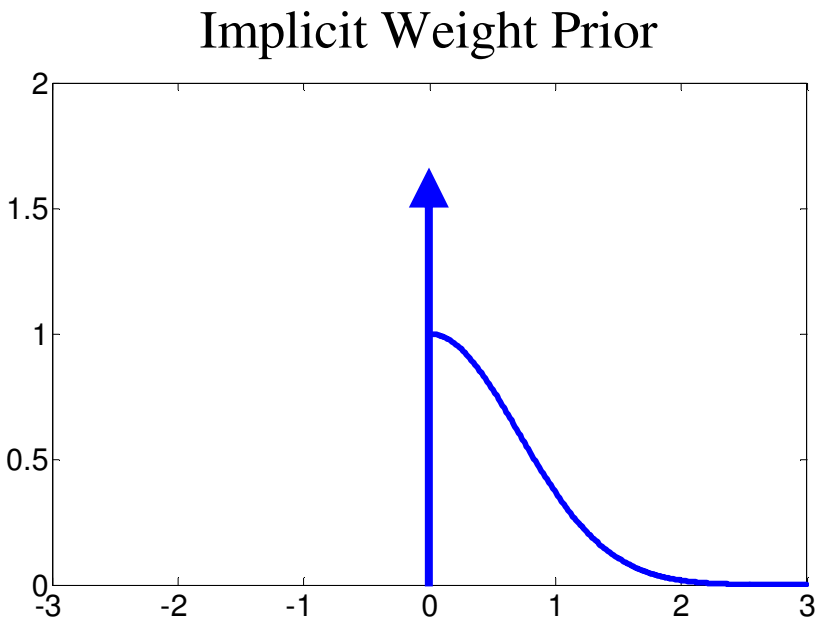
$$h\left(\boldsymbol{\gamma}^{-1}\right) = -\log\left|\lambda I + \Phi\Gamma\Phi^T\right|$$

- By working directly in $\mathbf{w}$ space, can easily add the constraint $\mathbf{w} \geq 0$.

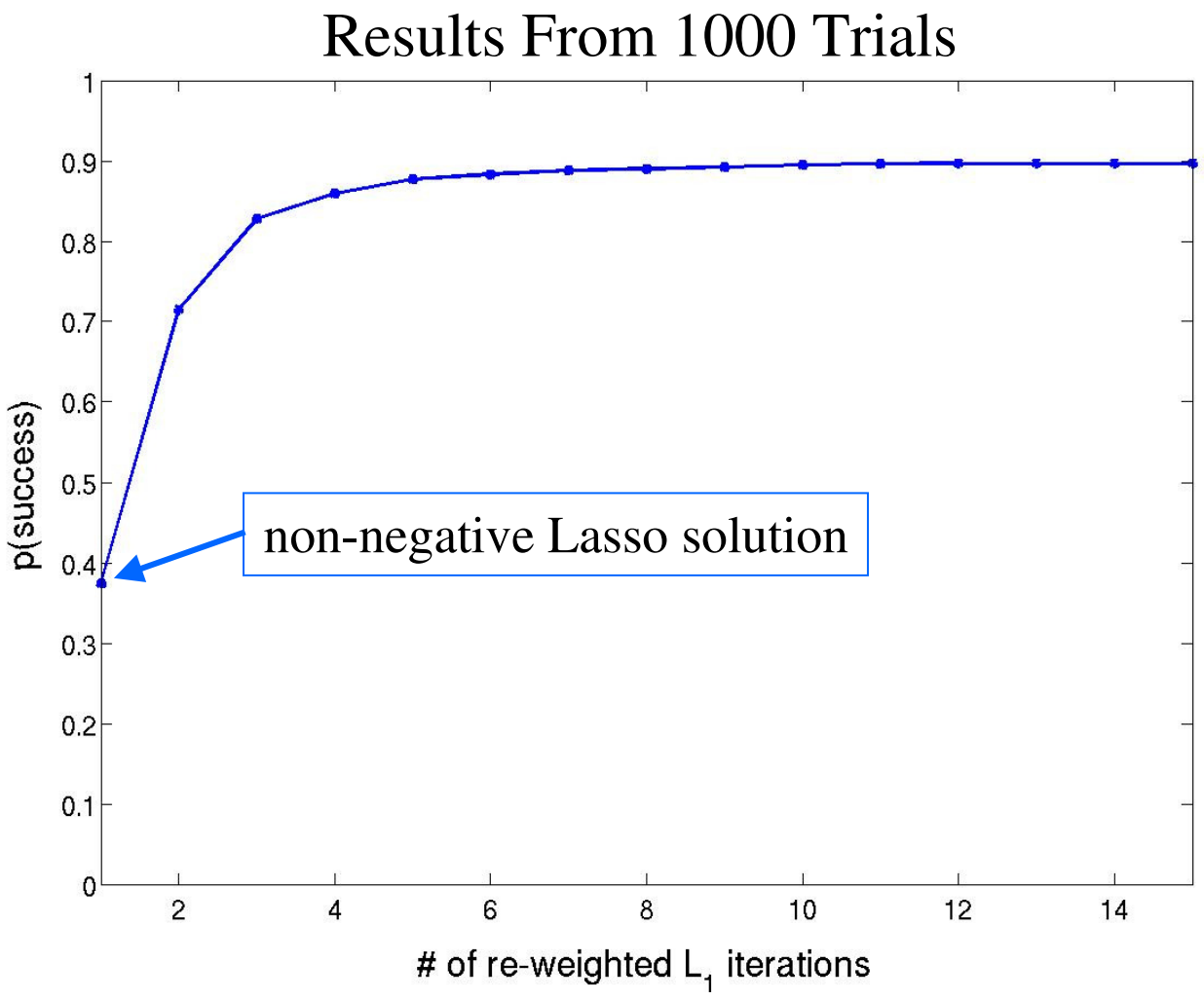
- Can solve using iterative re-weighted *non-negative* Lasso.

Empirical Example

- Generate data via $t = \Phi \mathbf{w}_0$:
  - Φ is 50 by 100 with Gaussian iid entries
  - $\mathbf{w}_0$ contains 30 random non-zero, non-negative entries.
- Run proposed algorithm and check after each iteration if $\mathbf{w}_0$ is recovered.



Empirical Example



Classification

- Current applications of the sparse linear model to classification require additional heuristic approximations.

- Can get around this using similar methods and iterative L_1 re-weighting.

Covariance Component Estimation

- We observe a data matrix T produced by the generative model

$$\mathrm{T} = \Phi\mathrm{W} + \mathrm{E}, \quad \mathrm{p}(\mathrm{W}) \propto \exp\left[-\tfrac{1}{2}\mathrm{W}^T\Sigma_w^{-1}\mathrm{W}\right], \quad \Sigma_w = \sum_i \gamma_i \mathrm{C}_i$$

- The resulting weights W and hyperparameters $\boldsymbol{\gamma}$ can be estimated using an analogous iterative re-weighted second-order cone (SOC) procedure.

- Applications:
  - MEG/EEG source localization
  - Compressed sensing
  - Multitask learning

Final Thoughts

- In the context of the sparse linear model, a variety of approximate inference methods can be reduced to hyperparameter MAP using some $p(\boldsymbol{\gamma}) \propto \exp[-f(\boldsymbol{\gamma})]$.

- Can do estimation and inference without conjugate hyperpriors; rather a large class of $f(\boldsymbol{\gamma})$ can be used.

- Choice should be application dependent and focus on the underlying cost function, not the plausibility of the associated full model.

- When comparing with EP, perhaps $f(\boldsymbol{\gamma})$ can be tailored to get similar Gaussian approximation in some cases. *Need not start with same full model to get the best results*.

Thank You