Learning (Convex Inference of Marginals)

Justin Domke

University of Maryland

Outline

Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

Experiments
- Introduction
- Results

Discussion
- Summary

Outline

Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

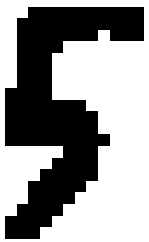
Experiments
- Introduction
- Results

Discussion
- Summary

Setting

y

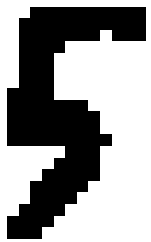
x



- True (unknown) distribution $p(\mathbf{x}, \mathbf{y})$
- Given $\mathbf{y}$, want to predict $p(x_i|\mathbf{y})$
- We have only $\{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\} \sim p(\mathbf{x}, \mathbf{y})$.

Setting

y **x**



- True (unknown) distribution $p(\mathbf{x}, \mathbf{y})$
- Given $\mathbf{y}$, want to predict $p(x_i|\mathbf{y})$
- We have only $\{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\} \sim p(\mathbf{x}, \mathbf{y})$.

Setting

- True (unknown) distribution $p(\mathbf{x}, \mathbf{y})$

Maximum Likelihood Approach:

1. Fit a graphical model $q(\mathbf{x}|\mathbf{y})$ by max (conditional) likelihood. (Learning)
2. Given $\mathbf{y}$, compute $q(x_i|\mathbf{y})$. (Inference)

Justification: Given a correct model, $q(\mathbf{x}|\mathbf{y}) \rightarrow p(\mathbf{x}|\mathbf{y})$.

Setting

- True (unknown) distribution $p(\mathbf{x}, \mathbf{y})$

Maximum Likelihood Approach:

1. Fit a graphical model $q(\mathbf{x}|\mathbf{y})$ by max (conditional) likelihood. (Learning)
2. Given $\mathbf{y}$, compute $q(x_i|\mathbf{y})$. (Inference)

Justification: Given a correct model, $q(\mathbf{x}|\mathbf{y}) \longrightarrow p(\mathbf{x}|\mathbf{y})$.

Problems with Max Likelihood

- In many cases, this works well.
- Problems arise, particularly when the model has high treewidth.
  - Computational Intractability
  - Model Defects

Problems with Max Likelihood

- In many cases, this works well.
- Problems arise, particularly when the model has high treewidth.
  - Computational Intractability
  - Model Defects

Problems with Max Likelihood

1. Computational Intractability.
   - Often, max likelihood can't be done.
   - Even if it could, would the results be what we want under approximate inference?

Problems with Max Likelihood

1. Computational Intractability.
   - Often, max likelihood can't be done.
   - Even if it could, would the results be what we want under approximate inference?

Problems with Max Likelihood

2. Model Defects
   - Possible to set θ such that $q(\mathbf{x}|\mathbf{y};\theta) = p(\mathbf{x}|\mathbf{y})$?
   - Max conditional likelihood $\approx$ E[KL-divergence]

$$\arg\min_\theta \sum_{\mathbf{y}} p(\mathbf{y}) \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log \frac{p(\mathbf{x}|\mathbf{y})}{q(\mathbf{x}|\mathbf{y};\theta)}$$

$$= \arg\min_\theta - \sum_{\mathbf{y}} p(\mathbf{y}) \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log q(\mathbf{x}|\mathbf{y};\theta)$$

$$\approx \arg\max_\theta \sum_{\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \log q(\hat{\mathbf{x}}|\hat{\mathbf{y}};\theta)$$

   - min KL-divergence $\neq$ best marginals.
     - (Even assuming exact inference.)

Problems with Max Likelihood

2. Model Defects

- Possible to set θ such that $q(\mathbf{x}|\mathbf{y};\theta) = p(\mathbf{x}|\mathbf{y})$?
- Max conditional likelihood $\approx$ E[KL-divergence]

$$\begin{aligned}
&\arg\min_\theta \sum_{\mathbf{y}} p(\mathbf{y}) \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log \frac{p(\mathbf{x}|\mathbf{y})}{q(\mathbf{x}|\mathbf{y};\theta)} \\
&= \arg\min_\theta - \sum_{\mathbf{y}} p(\mathbf{y}) \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log q(\mathbf{x}|\mathbf{y};\theta) \\
&\approx \arg\max_\theta \sum_{\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \log q(\hat{\mathbf{x}}|\hat{\mathbf{y}};\theta)
\end{aligned}$$

- min KL-divergence $\neq$ best marginals.
  - (Even assuming exact inference.)

Problems with Max Likelihood

2. Model Defects

- Possible to set θ such that $q(\mathbf{x}|\mathbf{y};\theta) = p(\mathbf{x}|\mathbf{y})$?
- Max conditional likelihood $\approx$ E[KL-divergence]

$$\begin{aligned}
&\arg\min_\theta \sum_{\mathbf{y}} p(\mathbf{y}) \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log \frac{p(\mathbf{x}|\mathbf{y})}{q(\mathbf{x}|\mathbf{y};\theta)} \\
&= \arg\min_\theta - \sum_{\mathbf{y}} p(\mathbf{y}) \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) \log q(\mathbf{x}|\mathbf{y};\theta) \\
&\approx \arg\max_\theta \sum_{\{(\hat{\mathbf{x}},\hat{\mathbf{y}})\}} \log q(\hat{\mathbf{x}}|\hat{\mathbf{y}};\theta)
\end{aligned}$$

- min KL-divergence $\neq$ best marginals.
  - (Even assuming exact inference.)

Outline

Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

Experiments
- Introduction
- Results

Discussion
- Summary

Overview

- **Goal**: Sidestep problems of intractability and model defects.
- In approximate inference, $q(\mathbf{x}|\mathbf{y})$ is often used to create a free energy function.

$$F(\mathbf{y}, \{b_r(\mathbf{x}_r)\}) \qquad \{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$$

- **Idea**: Think of F as mapping from $\mathbf{y}$ to $\{b_r^*(\mathbf{x}_r)\}$. Directly fit F to make the mapping as accurate as possible.
- Computational Tractability: Restrict F to be convex.
- Model Defects: Learn by minimizing empirical risk, where risk measures the accuracy of marginals.

Overview

- **Goal**: Sidestep problems of intractability and model defects.
- In approximate inference, $q(\mathbf{x}|\mathbf{y})$ is often used to create a free energy function.

$$F(\mathbf{y}, \{b_r(\mathbf{x}_r)\}) \qquad \{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$$

- **Idea**: Think of F as mapping from $\mathbf{y}$ to $\{b_r^*(\mathbf{x}_r)\}$. Directly fit F to make the mapping as accurate as possible.
- Computational Tractability: Restrict F to be convex.
- Model Defects: Learn by minimizing empirical risk, where risk measures the accuracy of marginals.

Overview

- **Goal**: Sidestep problems of intractability and model defects.
- In approximate inference, $q(\mathbf{x}|\mathbf{y})$ is often used to create a free energy function.

$$F(\mathbf{y}, \{b_r(\mathbf{x}_r)\}) \qquad \{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$$

- **Idea**: Think of F as mapping from $\mathbf{y}$ to $\{b_r^*(\mathbf{x}_r)\}$. Directly fit F to make the mapping as accurate as possible.
- Computational Tractability: Restrict F to be convex.
- Model Defects: Learn by minimizing empirical risk, where risk measures the accuracy of marginals.

Overview

- **Goal**: Sidestep problems of intractability and model defects.
- In approximate inference, $q(\mathbf{x}|\mathbf{y})$ is often used to create a free energy function.

$$F(\mathbf{y}, \{b_r(\mathbf{x}_r)\}) \qquad \{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$$

- **Idea**: Think of F as mapping from $\mathbf{y}$ to $\{b_r^*(\mathbf{x}_r)\}$. Directly fit F to make the mapping as accurate as possible.
- Computational Tractability: Restrict F to be convex.
- Model Defects: Learn by minimizing empirical risk, where risk measures the accuracy of marginals.

Overview

- **Goal**: Sidestep problems of intractability and model defects.
- In approximate inference, $q(\mathbf{x}|\mathbf{y})$ is often used to create a free energy function.

$$F(\mathbf{y}, \{b_r(\mathbf{x}_r)\}) \qquad \{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$$

- **Idea**: Think of F as mapping from $\mathbf{y}$ to $\{b_r^*(\mathbf{x}_r)\}$. Directly fit F to make the mapping as accurate as possible.
- Computational Tractability: Restrict F to be convex.
- Model Defects: Learn by minimizing empirical risk, where risk measures the accuracy of marginals.

Outline

Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

Experiments
- Introduction
- Results

Discussion
- Summary

Inference

$$F = \sum_{f \in \mathscr{F}} \sum_{r \in \mathscr{R}} \sum_{\mathbf{x}_r} w_f(\mathbf{x}_r, \mathbf{y}_r) f(b_r(\mathbf{x}_r))$$

- Typically, $\mathscr{F} = \{b, b \log b\}$. Anything convex over (0,1) is OK.

$$\{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$$

- Minimization is over some relaxation of the marginal polytope

$$\text{local consistency:} \quad \sum_{\mathbf{x}_{c \setminus i}} b_c(\mathbf{x}_c) = b_i(x_i)$$

$$\sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) = 1 \qquad \sum_{x_i} b_i(x_i) = 1$$

$$b_c(\mathbf{x}_c) \geq \mathbf{0} \qquad b_i(x_i) \geq 0$$

Inference

$$F = \sum_{f \in \mathscr{F}} \sum_{r \in \mathscr{R}} \sum_{\mathbf{x}_r} w_f(\mathbf{x}_r, \mathbf{y}_r) f(b_r(\mathbf{x}_r))$$

- Typically, $\mathscr{F} = \{b, b \log b\}$. Anything convex over (0,1) is OK.

$$\{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$$

- Minimization is over some relaxation of the marginal polytope

$$\text{local consistency:} \quad \sum_{\mathbf{x}_{c \setminus i}} b_c(\mathbf{x}_c) = b_i(x_i)$$

$$\sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) = 1 \qquad \sum_{x_i} b_i(x_i) = 1$$

$$b_c(\mathbf{x}_c) \geq \mathbf{0} \qquad b_i(x_i) \geq 0$$

Inference

$$F = \sum_{f \in \mathscr{F}} \sum_{r \in \mathscr{R}} \sum_{\mathbf{x}_r} w_f(\mathbf{x}_r, \mathbf{y}_r) f(b_r(\mathbf{x}_r))$$

- Typically, $\mathscr{F} = \{b, b \log b\}$. Anything convex over (0,1) is OK.

$$\{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$$

- Minimization is over some relaxation of the marginal polytope

$$\text{local consistency:} \quad \sum_{\mathbf{x}_{c \setminus i}} b_c(\mathbf{x}_c) = b_i(x_i)$$

$$\sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) = 1 \qquad \sum_{x_i} b_i(x_i) = 1$$

$$b_c(\mathbf{x}_c) \geq \mathbf{0} \qquad b_i(x_i) \geq 0$$

Inference

$$F = \sum_{f \in \mathscr{F}} \sum_{r \in \mathscr{R}} \sum_{\mathbf{x}_r} w_f(\mathbf{x}_r, \mathbf{y}_r) f(b_r(\mathbf{x}_r))$$

- Typically, $\mathscr{F} = \{b, b \log b\}$. Anything convex over (0,1) is OK.

$$\{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$$

- Minimization is over some relaxation of the marginal polytope

$$\text{local consistency:} \quad \sum_{\mathbf{x}_{c \setminus i}} b_c(\mathbf{x}_c) = b_i(x_i)$$

$$\sum_{\mathbf{x}_c} b_c(\mathbf{x}_c) = 1 \qquad \sum_{x_i} b_i(x_i) = 1$$

$$b_c(\mathbf{x}_c) \geq \mathbf{0} \qquad b_i(x_i) \geq 0$$

Inference

$$\{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} \sum_{f \in \mathscr{F}} \sum_{r \in \mathscr{R}} \sum_{\mathbf{x}_r} w_f(\mathbf{x}_r, \mathbf{y}_r) f(b_r(\mathbf{x}_r))$$

such that (local consistency)

Equivalent, more convenient formulation:

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathscr{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$

$$\mathbf{b} \leftrightarrow \{b_r(\mathbf{x}_r)\}, \forall r, \mathbf{x}_r$$

$$\mathbf{w}_f(\mathbf{y}) \leftrightarrow w_f(\mathbf{x}_r, \mathbf{y}_r), \forall r, \mathbf{x}_r$$

such that $A\mathbf{b} = \mathbf{d}$

$\mathbf{b} \geq \mathbf{0}$.

Inference

$$\{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} \sum_{f \in \mathscr{F}} \sum_{r \in \mathscr{R}} \sum_{\mathbf{x}_r} w_f(\mathbf{x}_r, \mathbf{y}_r) f(b_r(\mathbf{x}_r))$$

such that (local consistency)

Equivalent, more convenient formulation:

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathscr{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$

$$\mathbf{b} \leftrightarrow \{b_r(\mathbf{x}_r)\}, \forall r, \mathbf{x}_r$$

$$\mathbf{w}_f(\mathbf{y}) \leftrightarrow w_f(\mathbf{x}_r, \mathbf{y}_r), \forall r, \mathbf{x}_r$$

such that $A\mathbf{b} = \mathbf{d}$

$\mathbf{b} \geq \mathbf{0}$.

Inference

$$\{b_r^*(\mathbf{x}_r)\} = \arg\min_{\{b_r\}} \sum_{f \in \mathscr{F}} \sum_{r \in \mathscr{R}} \sum_{\mathbf{x}_r} w_f(\mathbf{x}_r, \mathbf{y}_r) f(b_r(\mathbf{x}_r))$$

such that (local consistency)

Equivalent, more convenient formulation:

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathscr{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$

such that $A\mathbf{b} = \mathbf{d}$

$\mathbf{b} \geq \mathbf{0}$.

$$\begin{aligned} \mathbf{b} &\leftrightarrow \{b_r(\mathbf{x}_r)\}, \forall r, \mathbf{x}_r \\ \mathbf{w}_f(\mathbf{y}) &\leftrightarrow w_f(\mathbf{x}_r, \mathbf{y}_r), \forall r, \mathbf{x}_r \end{aligned}$$

Outline

Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

Experiments
- Introduction
- Results

Discussion
- Summary

Learning

- Given $\{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\} \sim p(\mathbf{x}, \mathbf{y})$, how to quantify the quality of predicted marginals?
- Many possibilities. Two suggestions:
  - Log-loss
  - Quad-loss

Log-loss

- Try to minimize "expected average univariate KL-divergence".

$$
\begin{aligned}
F^* &= \arg\min_F \sum_{\mathbf{y}} p(\mathbf{y}) \sum_i \sum_{x_i} p(x_i|\mathbf{y}) \log \frac{p(x_i|\mathbf{y})}{b_i^*(x_i|\mathbf{y}, F)}. \\
&= \arg\min_F - \sum_{\mathbf{y}} p(\mathbf{y}) \sum_i \sum_{x_i} p(x_i|\mathbf{y}) \log b_i^*(x_i|\mathbf{y}, F) \\
&= \arg\min_F - \sum_{\mathbf{y}} \sum_i \sum_{x_i} p(x_i, \mathbf{y}) \log b_i^*(x_i|\mathbf{y}, F) \\
&\approx \arg\min_F \sum_{\{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\}} \underbrace{- \sum_i \log b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}_{L_{\log}}
\end{aligned}
$$

Log-loss

- Try to minimize "expected average univariate KL-divergence".

$$L_{\log} = -\sum_i \log b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)$$

$$\frac{\partial L_{\log}}{\partial \theta_j} = -\sum_i \frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)/\partial \theta_j}{b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}$$

(See also Kakade et al. "An Alternate Objective Function for Markovian Fields")

Quad-loss

- Try to minimize "expected average univariate quadratic difference".

$$
\begin{aligned}
F^* &= \arg\min_F \sum_{\mathbf{y}} p(\mathbf{y}) \sum_i \sum_{x_i} (p(x_i|\mathbf{y}) - b_i^*(x_i|\mathbf{y}, F))^2 \\
&= \arg\min_F \sum_{\mathbf{y}} \sum_i \sum_{x_i} \left(-2p(x_i, \mathbf{y}) b_i^*(x_i|\mathbf{y}, F) + p(\mathbf{y}) b_i^*(x_i|\mathbf{y}, F)^2\right) \\
&\approx \arg\min_F \sum_{\{(\hat{\mathbf{x}}, \hat{\mathbf{y}})\}} \underbrace{\sum_i \left(-2b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F) + \sum_{x_i} b_i^*(x_i|\hat{\mathbf{y}}, F)^2\right)}_{L_{\text{quad}}}
\end{aligned}
$$

Quad-loss

- Try to minimize "expected average univariate quadratic difference".

$$L_{\text{quad}} = \sum_i \left(-2b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F) + \sum_{x_i} b_i^*(x_i|\hat{\mathbf{y}}, F)^2\right)$$

$$\frac{\partial L_{\text{quad}}}{\partial \theta_j} = 2\sum_i \left(-\frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}{\partial \theta_j} + \sum_{x_i} b_i^*(x_i|\hat{\mathbf{y}}, F)\frac{\partial b_i^*(x_i|\hat{\mathbf{y}}, F)}{\partial \theta_j}\right)$$

Recap

$$\frac{\partial L_{\log}}{\partial \theta_j} = -\sum_i \frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)/\partial \theta_j}{b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}$$

$$\frac{\partial L_{\text{quad}}}{\partial \theta_j} = 2\sum_i \Big(-\frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}{\partial \theta_j} + \sum_{x_i} b_i^*(x_i|\hat{\mathbf{y}}, F)\frac{\partial b_i^*(x_i|\hat{\mathbf{y}}, F)}{\partial \theta_j} \Big)$$

If we could calculate $\frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}{\partial \theta_j}$, we could optimize $L_{\{\log,\text{quad}\}}$.
However, recall that $b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)$ is implicit.

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathscr{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$
$$\text{such that} \quad A\mathbf{b} = \mathbf{d}$$
$$\mathbf{b} \geq \mathbf{0}.$$

Recap

$$\frac{\partial L_{\text{log}}}{\partial \theta_j} = -\sum_i \frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)/\partial \theta_j}{b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}$$

$$\frac{\partial L_{\text{quad}}}{\partial \theta_j} = 2\sum_i \Big(- \frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}{\partial \theta_j} + \sum_{x_i} b_i^*(x_i|\hat{\mathbf{y}}, F) \frac{\partial b_i^*(x_i|\hat{\mathbf{y}}, F)}{\partial \theta_j} \Big)$$

If we could calculate $\frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}{\partial \theta_j}$, we could optimize $L_{\{\text{log,quad}\}}$.
However, recall that $b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)$ is implicit.

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathcal{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$

$$\text{such that} \quad A\mathbf{b} = \mathbf{d}$$

$$\mathbf{b} \geq \mathbf{0}.$$

Recap

$$\frac{\partial L_{\text{log}}}{\partial \theta_j} = -\sum_i \frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)/\partial \theta_j}{b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}$$

$$\frac{\partial L_{\text{quad}}}{\partial \theta_j} = 2\sum_i \Big(- \frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}{\partial \theta_j} + \sum_{x_i} b_i^*(x_i|\hat{\mathbf{y}}, F) \frac{\partial b_i^*(x_i|\hat{\mathbf{y}}, F)}{\partial \theta_j} \Big)$$

If we could calculate $\frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}{\partial \theta_j}$, we could optimize $L_{\{\text{log,quad}\}}$.

However, recall that $b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)$ is implicit.

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathcal{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$

$$\text{such that} \quad A\mathbf{b} = \mathbf{d}$$

$$\mathbf{b} \geq \mathbf{0}.$$

Recap

$$\frac{\partial L_{\text{log}}}{\partial \theta_j} = -\sum_i \frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)/\partial \theta_j}{b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}$$

$$\frac{\partial L_{\text{quad}}}{\partial \theta_j} = 2\sum_i \Big(-\frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}{\partial \theta_j} + \sum_{x_i} b_i^*(x_i|\hat{\mathbf{y}}, F)\frac{\partial b_i^*(x_i|\hat{\mathbf{y}}, F)}{\partial \theta_j}\Big)$$

If we could calculate $\frac{\partial b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)}{\partial \theta_j}$, we could optimize $L_{\{\text{log,quad}\}}$.
However, recall that $b_i^*(\hat{x}_i|\hat{\mathbf{y}}, F)$ is <u>implicit</u>.

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathscr{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$
$$\text{such that} \quad A\mathbf{b} = \mathbf{d}$$
$$\mathbf{b} \geq \mathbf{0}.$$

Outline

Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

Experiments
- Introduction
- Results

Discussion
- Summary

Claim 1

Claim 1: Let $F(\mathbf{b}, \theta)$ be a continuous function such that for all θ, F that has a unique stationary point in $\mathbf{b}$.Define $\mathbf{b}^*(\theta)$ such that $\frac{\partial F(\mathbf{b}^*(\theta),\theta)}{\partial \mathbf{b}} = \mathbf{0}$.Then,

$$\frac{\partial \mathbf{b}^*(\theta)}{\partial \theta_j} = -\Big(\frac{\partial^2 F(\mathbf{b}^*(\theta),\theta)}{\partial \mathbf{b} \partial \mathbf{b}^T}\Big)^{-1} \frac{\partial^2 F(\mathbf{b}^*(\theta),\theta)}{\partial \mathbf{b} \partial \theta_j}.$$

Proof: (Implicit Function Theorem.)

Claim 1

Claim 1: Let $F(\mathbf{b}, \theta)$ be a continuous function such that for all θ, F that has a unique stationary point in $\mathbf{b}$.Define $\mathbf{b}^*(\theta)$ such that $\frac{\partial F(\mathbf{b}^*(\theta),\theta)}{\partial \mathbf{b}} = \mathbf{0}$.Then,

$$\frac{\partial \mathbf{b}^*(\theta)}{\partial \theta_j} = -\Big(\frac{\partial^2 F(\mathbf{b}^*(\theta),\theta)}{\partial \mathbf{b} \partial \mathbf{b}^T}\Big)^{-1} \frac{\partial^2 F(\mathbf{b}^*(\theta),\theta)}{\partial \mathbf{b} \partial \theta_j}.$$

- Not good enough, since F is minimized under constraints.

Claim 2

Claim 2: Define $\mathbf{b}^*(\theta) \doteq \arg\min_{\mathbf{b}} F(\mathbf{b}, \theta)$, such that $A\mathbf{b} = d$ for some convex function F. Then,

$$\frac{\partial \mathbf{b}^*(\theta)}{\partial \theta_j} = (D^{-1}A^T(AD^{-1}A^T)^{-1}AD^{-1} - D^{-1})\frac{\partial^2 F}{\partial \mathbf{b} \partial \theta_j},\ D = (\frac{\partial^2 F}{\partial \mathbf{b} \partial \mathbf{b}^T}).$$

Proof: (Make a Lagrangian, apply claim 1, do algebra.)

Claim 2

Claim 2: Define $\mathbf{b}^*(\theta) \doteq \arg\min_{\mathbf{b}} F(\mathbf{b}, \theta)$, such that $A\mathbf{b} = d$ for some convex function F. Then,

$$\frac{\partial \mathbf{b}^*(\theta)}{\partial \theta_j} = (D^{-1}A^T(AD^{-1}A^T)^{-1}AD^{-1} - D^{-1})\frac{\partial^2 F}{\partial \mathbf{b}\partial \theta_j},\ D = (\frac{\partial^2 F}{\partial \mathbf{b}\partial \mathbf{b}^T}).$$

In our case,

$$D = \mathrm{diag}(\sum_f \mathbf{w}_f(\mathbf{y}) \odot f''(\mathbf{b}))$$

$$\frac{\partial^2 F}{\partial \mathbf{b}\partial \theta_j} = \sum_f \frac{\partial \mathbf{w}_f(\mathbf{y})}{\partial \theta_j} \odot f'$$

We still need $\frac{\partial \mathbf{w}_f(\mathbf{y})}{\partial \theta_j}$.

Claim 2

Claim 2: Define $\mathbf{b}^*(\theta) \doteq \arg\min_{\mathbf{b}} F(\mathbf{b}, \theta)$, such that $A\mathbf{b} = d$ for some convex function F. Then,

$$\frac{\partial \mathbf{b}^*(\theta)}{\partial \theta_j} = (D^{-1}A^T(AD^{-1}A^T)^{-1}AD^{-1} - D^{-1})\frac{\partial^2 F}{\partial \mathbf{b}\partial \theta_j},\ D = (\frac{\partial^2 F}{\partial \mathbf{b}\partial \mathbf{b}^T}).$$

In our case,

$$D = \mathrm{diag}(\sum_f \mathbf{w}_f(\mathbf{y}) \odot f''(\mathbf{b}))$$

$$\frac{\partial^2 F}{\partial \mathbf{b}\partial \theta_j} = \sum_f \frac{\partial \mathbf{w}_f(\mathbf{y})}{\partial \theta_j} \odot f'$$

We still need $\frac{\partial \mathbf{w}_f(\mathbf{y})}{\partial \theta_j}$.

Derivatives of beliefs- the bottom line

- Want to calculate $\frac{\partial L}{\partial \theta_j}$. Procedure:
  1. Run some optimization

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathcal{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$
$$\text{such that} \quad A\mathbf{b} = \mathbf{d}$$
$$\mathbf{b} \geq \mathbf{0}.$$

  2. Solve the linear system given by Claim 2 to get $\frac{\partial L}{\partial \theta_j}$.

- When learning, <u>two</u> optimizations:
  1. "inner" optimization (over $\{b_r\}$)
  2. "outer" optimization (over θ)

Derivatives of beliefs- the bottom line

- Want to calculate $\frac{\partial L}{\partial \theta_j}$. Procedure:
  1. Run some optimization

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathcal{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$
$$\text{such that} \quad A\mathbf{b} = \mathbf{d}$$
$$\mathbf{b} \geq \mathbf{0}.$$

  2. Solve the linear system given by Claim 2 to get $\frac{\partial L}{\partial \theta_j}$.

- When learning, two optimizations:
  1. "inner" optimization (over $\{b_r\}$)
  2. "outer" optimization (over θ)

Derivatives of beliefs- the bottom line

- Want to calculate $\frac{\partial L}{\partial \theta_j}$. Procedure:
  1. Run some optimization

$$\mathbf{b}^* = \arg\min_{\mathbf{b}} \sum_{f \in \mathcal{F}} \mathbf{w}_f(\mathbf{y})^T f(\mathbf{b})$$
$$\text{such that} \quad A\mathbf{b} = \mathbf{d}$$
$$\mathbf{b} \geq \mathbf{0}.$$

  2. Solve the linear system given by Claim 2 to get $\frac{\partial L}{\partial \theta_j}$.

- When learning, <u>two</u> optimizations:
  1. "inner" optimization (over $\{b_r\}$)
  2. "outer" optimization (over θ)

Outline

Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

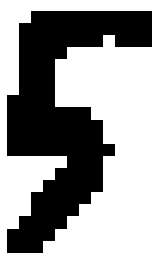
Experiments
- Introduction
- Results

Discussion
- Summary

Problem

- "Denoising" of 10 binary images of each class (1-9) from the MNIST database.
  - **y** is the observed, noisy image

  - **x** is the unobserved, clean image



Setup

- Use regions consisting of individual variables, and neighboring pairs.
- Total of 40 parameters. (Somewhat redundant)
  - $w_b(\mathbf{x}_c, \mathbf{y}_c)$ (initialize to 0)
  - $w_b(\mathbf{x}_i, \mathbf{y}_i)$ (initialize to 0)
  - $w_{b \log b}(\mathbf{x}_c, \mathbf{y}_c)$ (initialize to 1)
  - $w_{b \log b}(\mathbf{x}_i, \mathbf{y}_i)$ (initialize to 1)
- "Inner loop" optimization uses PDCO interior method.
  - Tolerances very strict.
- "Outer loop" optimization uses Matlab's BFGS.

Setup

- Use regions consisting of individual variables, and neighboring pairs.
- Total of 40 parameters. (Somewhat redundant)
  - $w_b(\mathbf{x}_c, \mathbf{y}_c)$ (initialize to 0)
  - $w_b(\mathbf{x}_i, \mathbf{y}_i)$ (initialize to 0)
  - $w_{b \log b}(\mathbf{x}_c, \mathbf{y}_c)$ (initialize to 1)
  - $w_{b \log b}(\mathbf{x}_i, \mathbf{y}_i)$ (initialize to 1)
- "Inner loop" optimization uses PDCO interior method.
  - Tolerances very strict.
- "Outer loop" optimization uses Matlab's BFGS.

Setup

- Use regions consisting of individual variables, and neighboring pairs.
- Total of 40 parameters. (Somewhat redundant)
  - $w_b(\mathbf{x}_c, \mathbf{y}_c)$ (initialize to 0)
  - $w_b(\mathbf{x}_i, \mathbf{y}_i)$ (initialize to 0)
  - $w_{b \log b}(\mathbf{x}_c, \mathbf{y}_c)$ (initialize to 1)
  - $w_{b \log b}(\mathbf{x}_i, \mathbf{y}_i)$ (initialize to 1)
- "Inner loop" optimization uses PDCO interior method.
  - Tolerances very strict.
- "Outer loop" optimization uses Matlab's BFGS.

Setup

- Use regions consisting of individual variables, and neighboring pairs.
- Total of 40 parameters. (Somewhat redundant)
  - $w_b(\mathbf{x}_c, \mathbf{y}_c)$ (initialize to 0)
  - $w_b(\mathbf{x}_i, \mathbf{y}_i)$ (initialize to 0)
  - $w_{b \log b}(\mathbf{x}_c, \mathbf{y}_c)$ (initialize to 1)
  - $w_{b \log b}(\mathbf{x}_i, \mathbf{y}_i)$ (initialize to 1)
- "Inner loop" optimization uses PDCO interior method.
  - Tolerances very strict.
- "Outer loop" optimization uses Matlab's BFGS.

Comparison

- Compare to CRF toolbox.
  - Vishwanathan et al., ICML 2006
  - `www.cs.ubc.ca/~murphyk/Software/CRF/crf.html`
- Inference: mean-field or (loopy) belief propagation.
- Learning: Pseudolikelihood, or surrogate to likelihood where inference algorithm is used to approximate marginals defining the gradient.
- Features are constant + indicator for each possible configuration of each variable/pair.

Outline

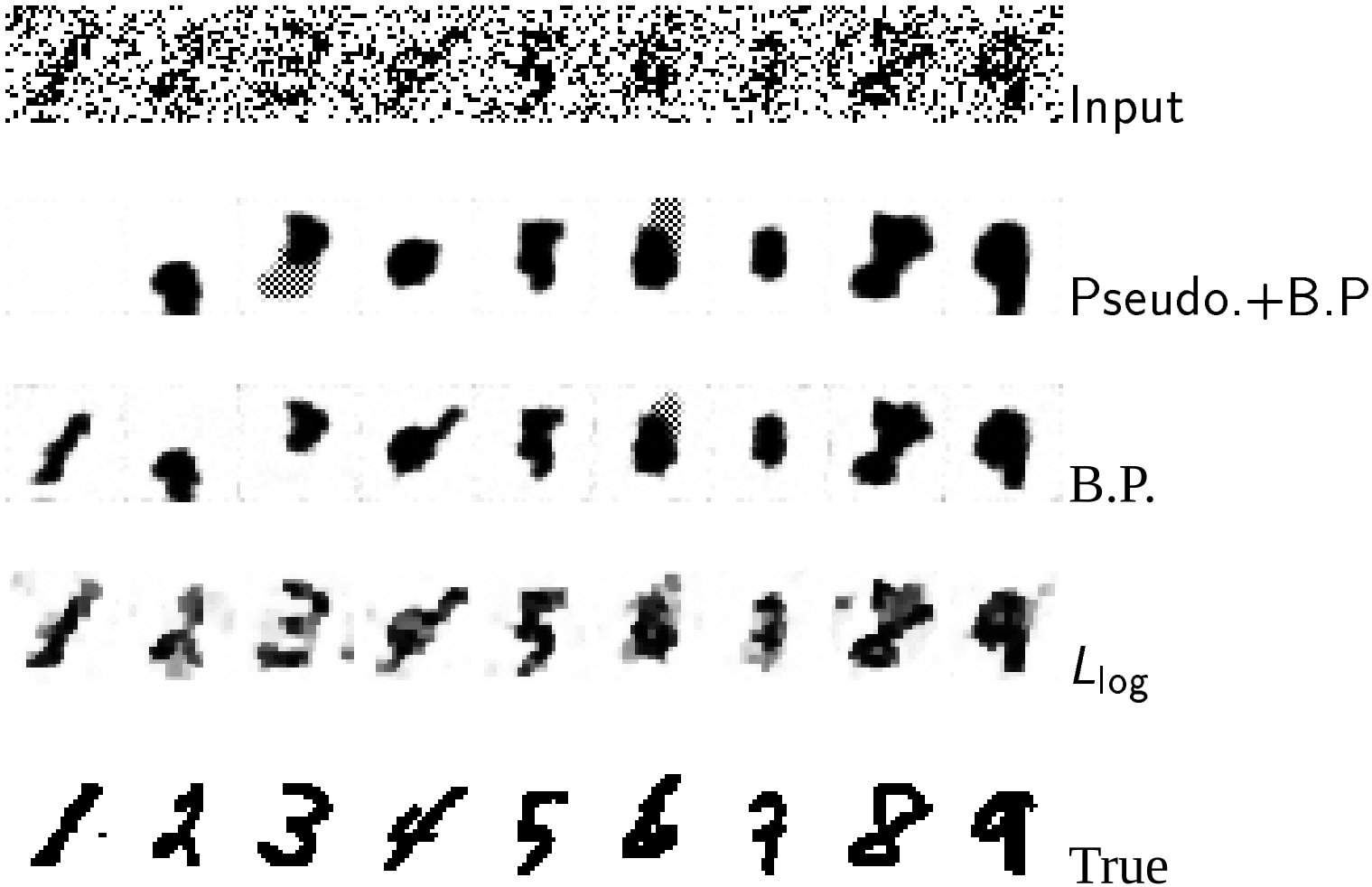
Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

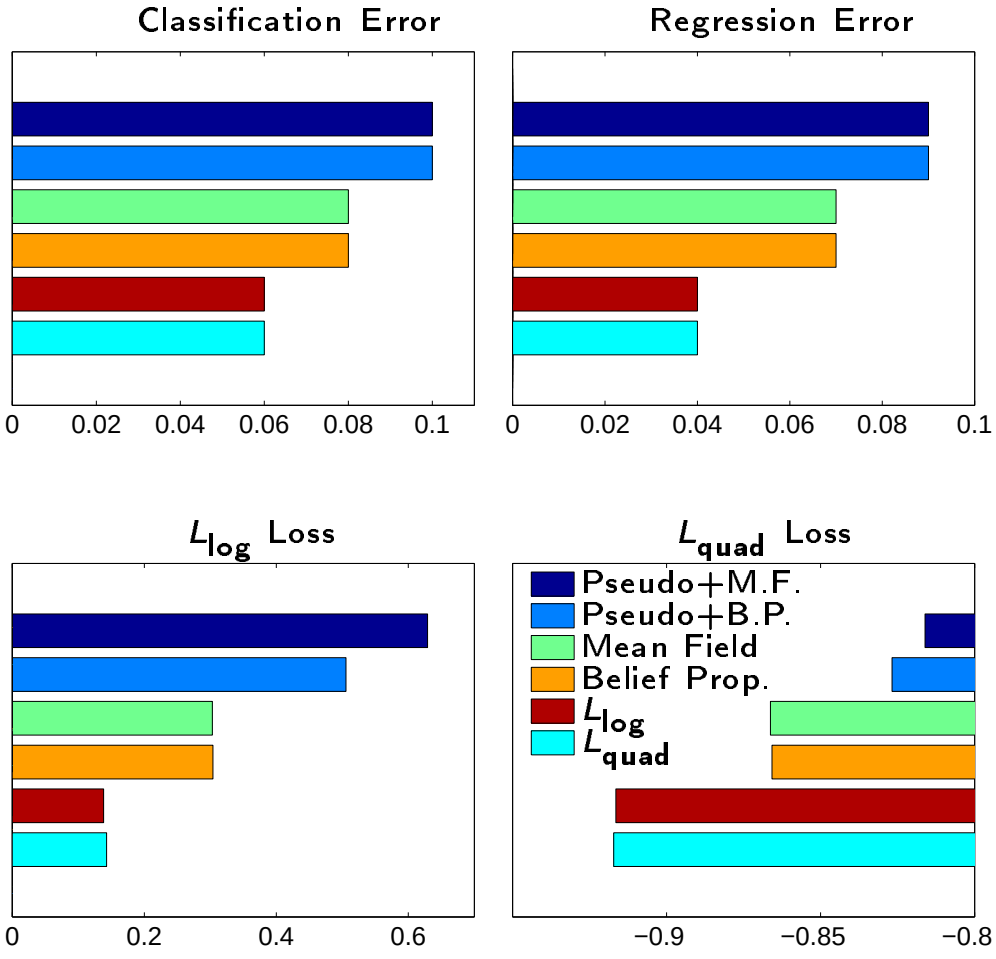
Experiments
- Introduction
- Results

Discussion
- Summary

50% noise



50% noise


Classification Error
Regression Error
L_{log} Loss
L_{quad} Loss
Pseudo+M.F.
Pseudo+B.P.
Mean Field
Belief Prop.
L_{log}
L_{quad}
0
0.02
0.04
0.06
0.08
0.1
0.2
0.4
0.6
−0.9
−0.85
−0.8

Outline

Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

Experiments
- Introduction
- Results

Discussion
- Summary

Summary

- Common to fit classifiers by minimizing a loss close to classification error.
- Graphical models, however, usually use a score (the likelihood) that is both
  - Difficult to optimize.
  - Remote from empirical risk.
- This paper presents an approach for "fitting a free energy" to directly give good marginals.

Summary

- Common to fit classifiers by minimizing a loss close to classification error.
- Graphical models, however, usually use a score (the likelihood) that is both
  - Difficult to optimize.
  - Remote from empirical risk.
- This paper presents an approach for "fitting a free energy" to directly give good marginals.

Summary

- Common to fit classifiers by minimizing a loss close to classification error.
- Graphical models, however, usually use a score (the likelihood) that is both
  - Difficult to optimize.
  - Remote from empirical risk.
- This paper presents an approach for "fitting a free energy" to directly give good marginals.

The End

Thank you.

Future Work:

1. Different loss functions.
2. Better entropy term.
3. Exploit hidden variables.
4. Better relaxation of marginal polytope.
5. Better "inner loop" optimization.
6. Better F.

Outline

Introduction
- Motivation
- Overview of the approach

Approach
- Inference
- Learning (Loss functions)
- Learning (Derivatives of beliefs)

Experiments
- Introduction
- Results

Discussion
- Summary

.

Future Work

1. Different loss functions.
   - For example, common to find $\arg\max_{x_i} b^*(x_i|\mathbf{y})$.
   - Could use a loss function (approximating) the univariate classification error.

(Gross et al. "Training CRFs for maximum labelwise accuracy", NIPS 2006)

Future Work

2. More flexible entropy term.
   - As described, this approach requires fixing $w_{b \log b} > 0$ to guarantee convexity.
   - However, negative entropy terms can be allowed while preserving convexity over the locally consistent marginals.
   - Easiest approach- just pick a better (fixed) entropy approximation.

(Heskes, "Convexity Arguments for Efficient Minimization of the Bethe and Kikuchi Free Energies", JAIR 2006)

Entropy

$w_{b\log b}(\mathbf{x}_c, \mathbf{y}_c)$ and $w_{b\log b}(x_i, y_i)$ for L_{quad} with 50% noise

$\mathbf{x}_c \backslash \mathbf{y}_c$	(0,0)	(0,1)	(1,0)	(1,1)
(0,0)	4.86	0.04	0.05	0.02
(0,1)	4.22	3.86	4.54	5.00
(1,0)	4.13	4.49	2.14	5.13
(1,1)	0.06	0.02	0.03	0.02

$x_i \backslash y_i$	0	1
0	4.47	0.02
1	0.03	0.03

Future Work

3. Hidden variables.
   - The same algorithm can be used with hidden variables, by taking the sum over the variables in $L_{\{\log,\text{quad}\}}$ over the observed variables.

Future Work

4. Better relaxation of the marginal polytope.
   - Currently, the model must try to compensate during the learning stage for defects in the marginal polytope.

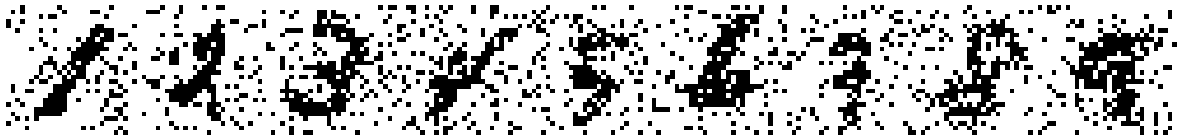
Future Work

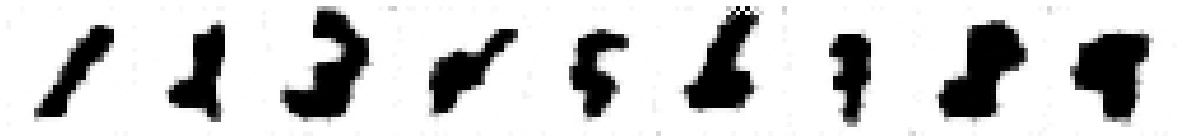
5. Better "inner loop" optimization.
   - Generic optimization is OK for 28x28 images.
   - Derive a message passing algorithm?

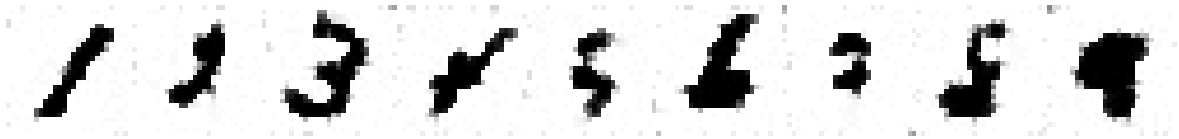
Future Work

6. More general function F.
   - Any function $F(\mathbf{y}, \{b_r(\mathbf{x}_r)\})$ can give an implicit mapping.
     - As long as it is convex and continuous, learning should be doable by implicit differentiation.
   - A larger set $\mathscr{F}$ than $\{b, b \log b\}$?
   - Some F based on different principles?

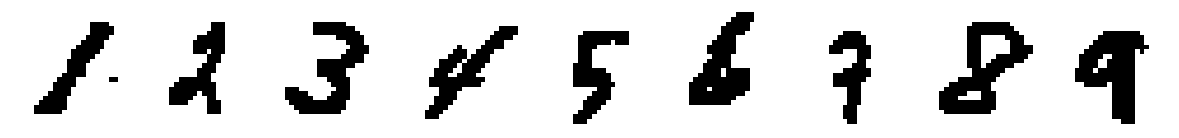
30% noise


Input

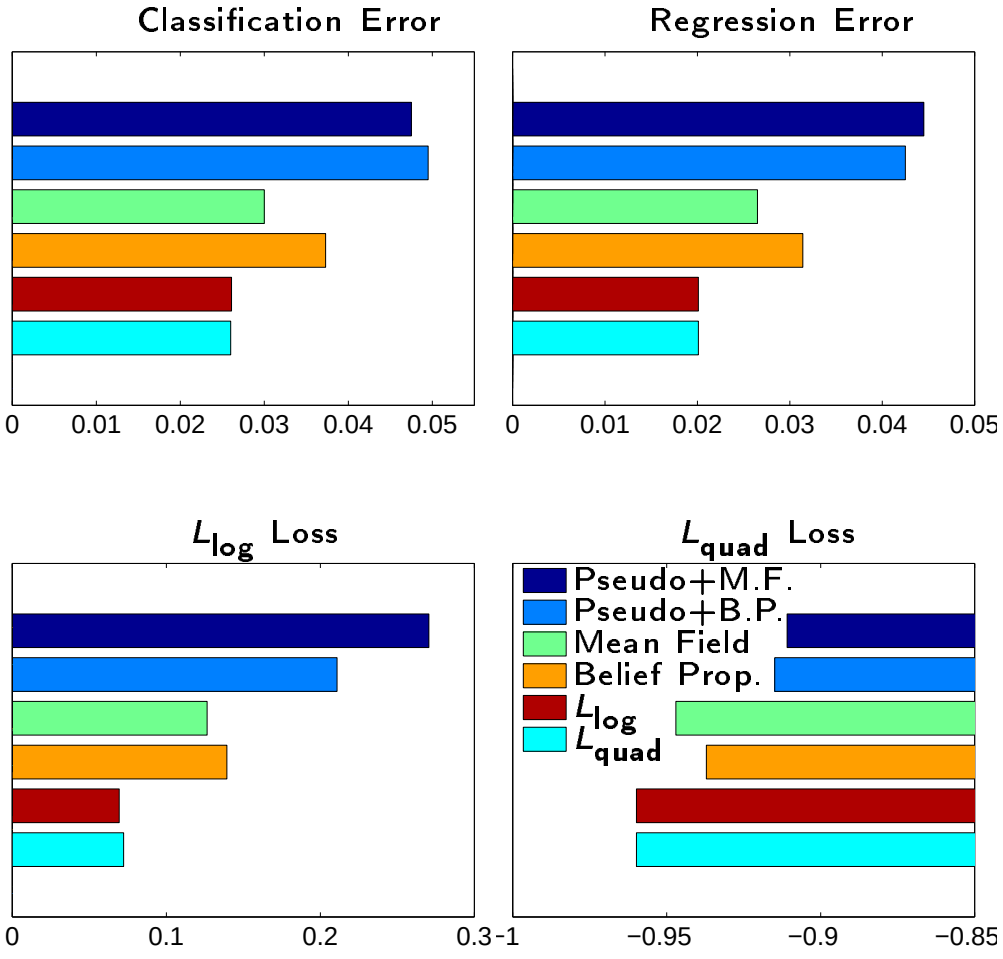

Pseudo.+B.P


B.P.


$L_{\log}$


True

30% noise


Classification Error
Regression Error
L_{log} Loss
L_{quad} Loss
Pseudo+M.F.
Pseudo+B.P.
Mean Field
Belief Prop.
L_{log}
L_{quad}
0
0.01
0.02
0.03
0.04
0.05
0.1
0.2
0.3
-1
−0.95
−0.9
−0.85