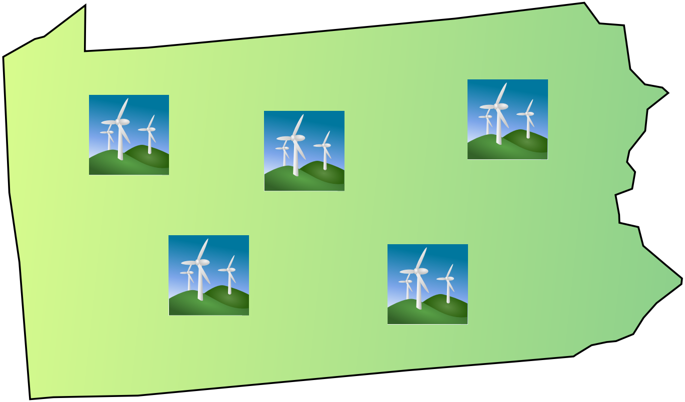
Sparse Gaussian Conditional Random Fields: Algorithms, Theory, and Application to Energy Forecasting

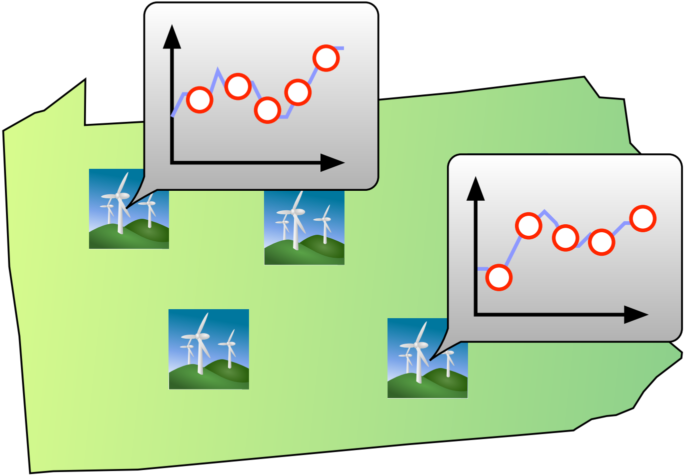
Matt Wytock, Zico Kolter
Carnegie Mellon University

November 11, 2013

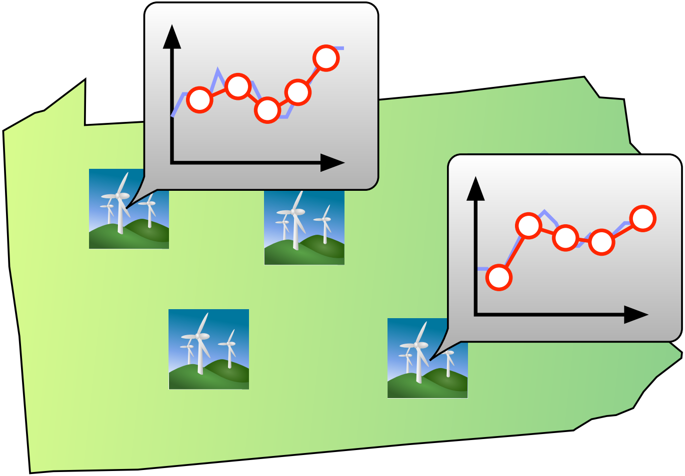
Wind power forecasting



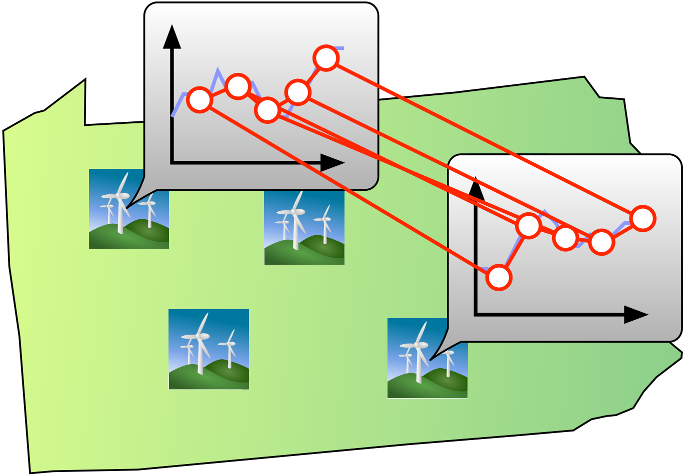
Wind power forecasting



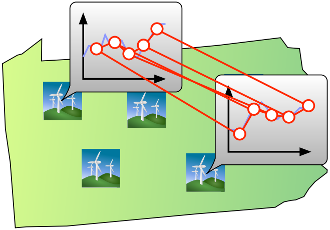
Wind power forecasting



Wind power forecasting



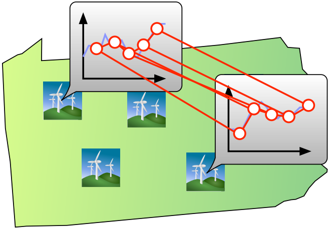
The sparse Gaussian CRF model



Outputs: wind power, $y \in \mathbb{R}^p$

$$y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_p$$

The sparse Gaussian CRF model



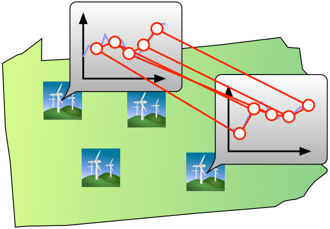
Outputs: wind power, $y \in \mathbb{R}^p$

$$y_1 \quad y_2 \quad y_3 \quad \cdots \quad y_p$$

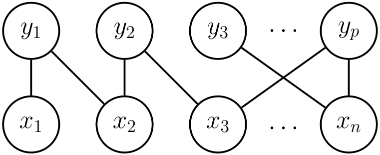
$$x_1 \quad x_2 \quad x_3 \quad \cdots \quad x_n$$

Inputs: past wind power and weather forecasts, $x \in \mathbb{R}^n$

The sparse Gaussian CRF model

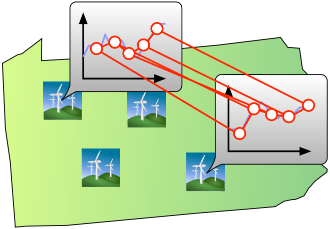


Outputs: wind power, $y \in \mathbb{R}^p$

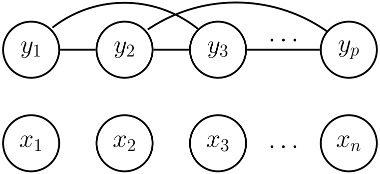


Inputs: past wind power and weather forecasts, $x \in \mathbb{R}^n$

The sparse Gaussian CRF model

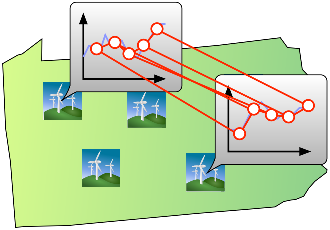


Outputs: wind power, $y \in \mathbb{R}^p$

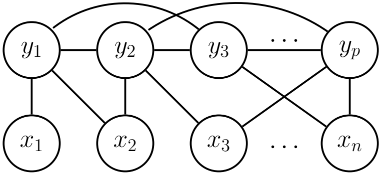


Inputs: past wind power and weather forecasts, $x \in \mathbb{R}^n$

The sparse Gaussian CRF model



Outputs: wind power, $y \in \mathbb{R}^p$



Inputs: past wind power and weather forecasts, $x \in \mathbb{R}^n$

Sparse regression and sparse inverse covariance estimation

- ℓ_1 methods very popular for high-dimensional regression and estimating high-dimensional undirected graphical models (Gaussian MRF)

- Sohn and Kim (2012) and Yuan and Zhang (2012) also independently propose the sparse Gaussian CRF model and consider applications to computational biology, computer vision, natural language processing and finance

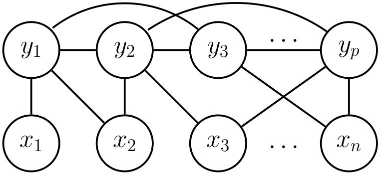
Contributions

- Second-order active set algorithm several orders of magnitude faster than previously used algorithms

- Theoretical analysis with bounds depending logarithmically on the data dimension and polynomially on max degree of the CRF

- State-of-the-art performance on two large-scale energy forecasting problems

Optimization problem



- We model the conditional distribution

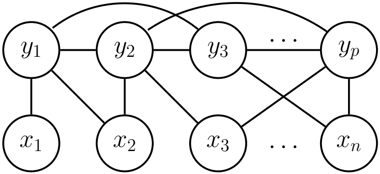
$$p(y|x) \propto \exp\left(-y^T \Lambda y - 2x^T \Theta y\right)$$

- Maximum likelihood estimation with ℓ_1 regularization

$$\begin{aligned}\underset{\Lambda,\Theta}{\text{minimize}} \; & -\log|\Lambda| + \operatorname{tr}\Lambda S_{yy} + 2\operatorname{tr}\Theta S_{yx} + \operatorname{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta \\ & + \lambda\|\Lambda\|_1 + \lambda\|\Theta\|_1\end{aligned}$$

- Convex but difficult to optimize due to matrix fractional term

Optimization problem



- We model the conditional distribution

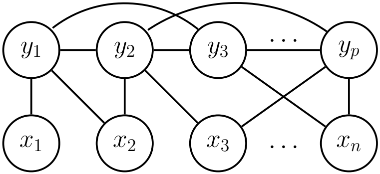
$$p(y|x) \propto \exp\left(-y^T \Lambda y - 2x^T \Theta y\right)$$

- Maximum likelihood estimation with ℓ_1 regularization

$$\begin{aligned}\underset{\Lambda,\Theta}{\text{minimize}} & -\log|\Lambda| + \operatorname{tr}\Lambda S_{yy} + 2\operatorname{tr}\Theta S_{yx} + \operatorname{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta \\ & + \lambda\|\Lambda\|_1 + \lambda\|\Theta\|_1\end{aligned}$$

- Convex but difficult to optimize due to matrix fractional term

Optimization problem



- We model the conditional distribution

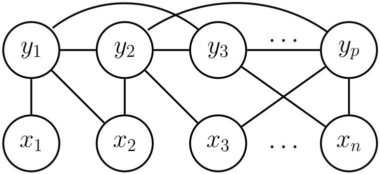
$$p(y|x) \propto \exp\left(-y^T \Lambda y - 2x^T \Theta y\right)$$

- Maximum likelihood estimation with ℓ_1 regularization

$$\begin{aligned}\underset{\Lambda,\Theta}{\text{minimize}} - \log|\Lambda| + \operatorname{tr}\Lambda S_{yy} + 2\operatorname{tr}\Theta S_{yx} + \operatorname{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta \\ + \lambda\|\Lambda\|_1 + \lambda\|\Theta\|_1\end{aligned}$$

- Convex but difficult to optimize due to matrix fractional term

Optimization problem



- We model the conditional distribution

$$p(y|x) \propto \exp\left(-y^T \Lambda y - 2x^T \Theta y\right)$$

- Maximum likelihood estimation with ℓ_1 regularization

$$\begin{aligned} \underset{\Lambda,\Theta}{\text{minimize}} & -\log|\Lambda| + \operatorname{tr}\Lambda S_{yy} + 2\operatorname{tr}\Theta S_{yx} + \operatorname{tr}\Lambda^{-1}\Theta^T S_{xx}\Theta \\ & + \lambda\|\Lambda\|_1 + \lambda\|\Theta\|_1 \end{aligned}$$

- Convex but difficult to optimize due to matrix fractional term

Second-order active set method

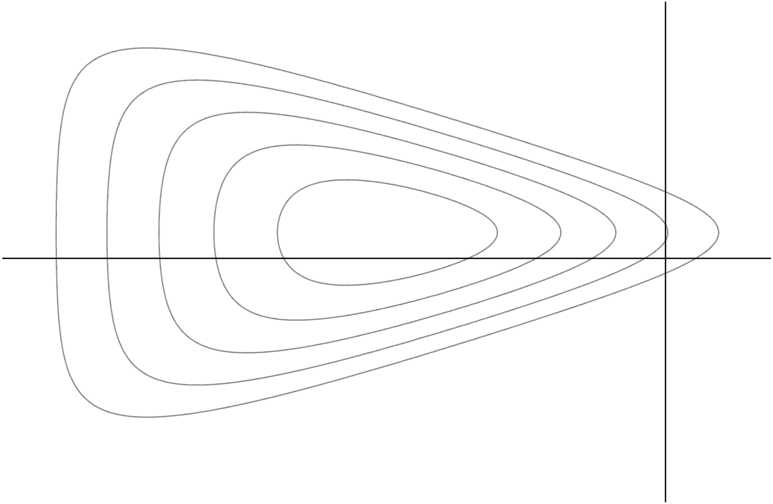
- We develop a second-order method using the framework defined by Tseng and Yun (2009) and Hsieh et al. (2011)

- while not converged
  1. Form the second-order Taylor expansion
  $$\hat{f}(x+\Delta) = f(x) + \nabla_x f(x)^T \Delta + \frac{1}{2}\Delta^T \nabla_x^2 f(x) \Delta$$
  2. Solve for the regularized Newton step
  $$d = \arg\min_{\Delta} \hat{f}(x+\Delta) + \lambda \|x+\Delta\|_1$$
  3. Update x using backtracking line search

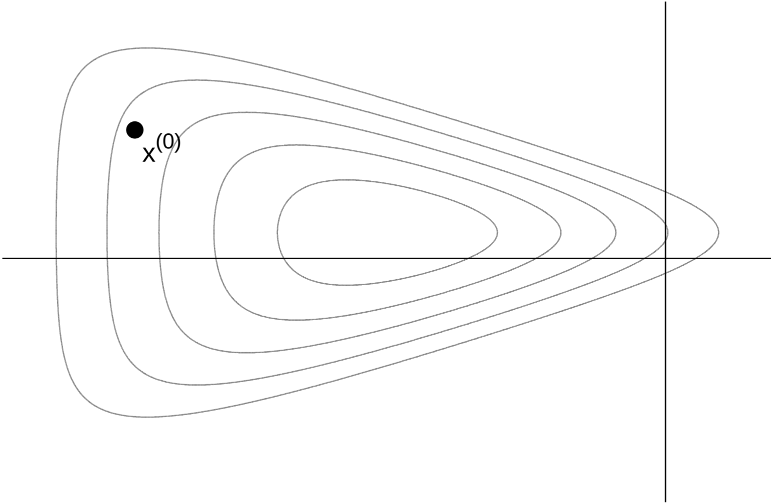
- Newton step cannot be found in closed form so we use coordinate descent with an active set

- Other performance tricks, Matlab/C++ version available

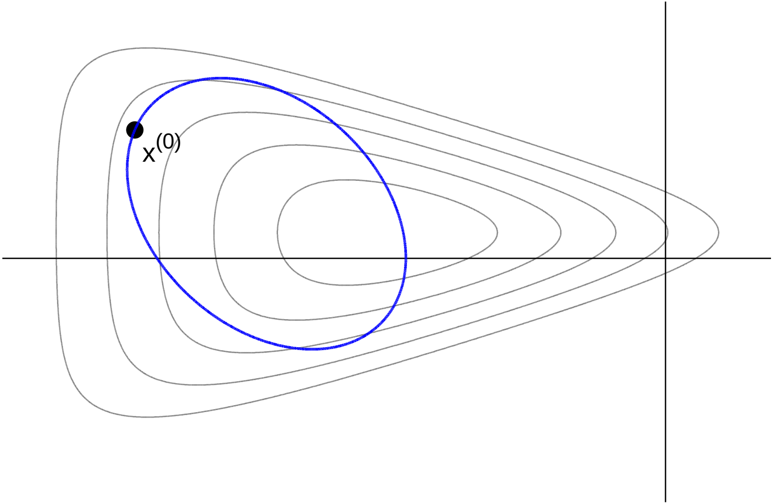
Optimization example



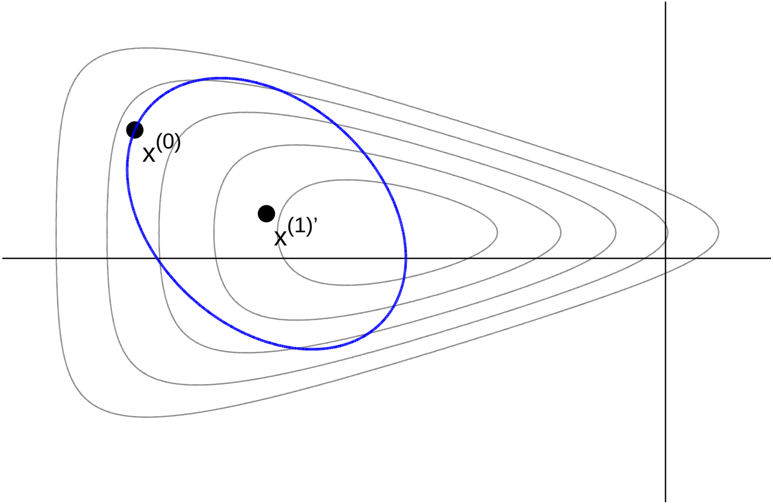
Optimization example



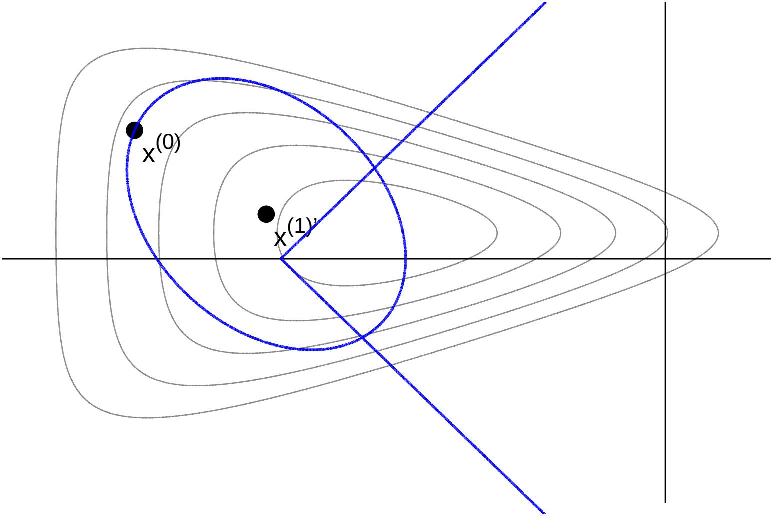
Optimization example



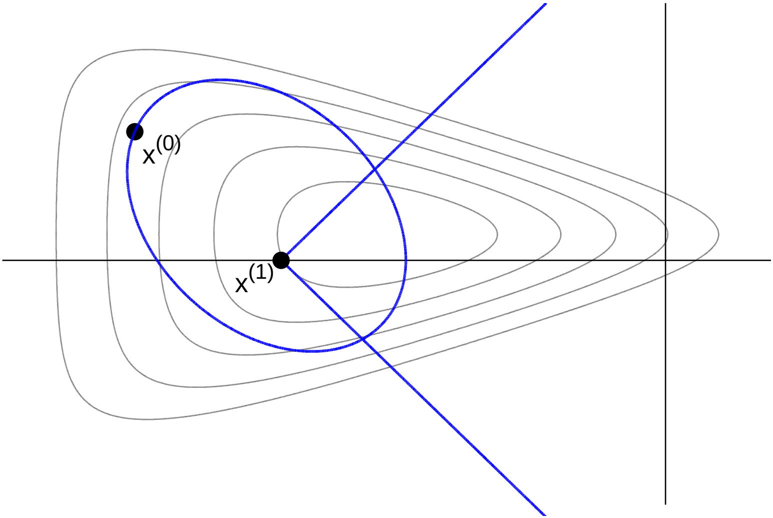
Optimization example



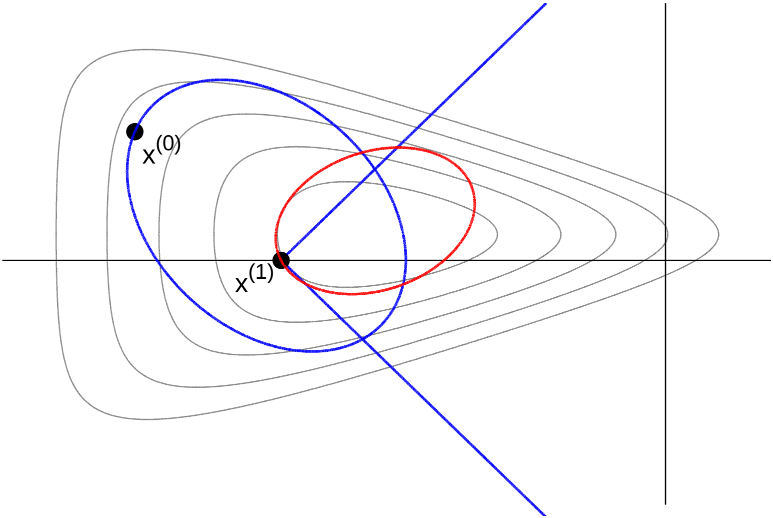
Optimization example



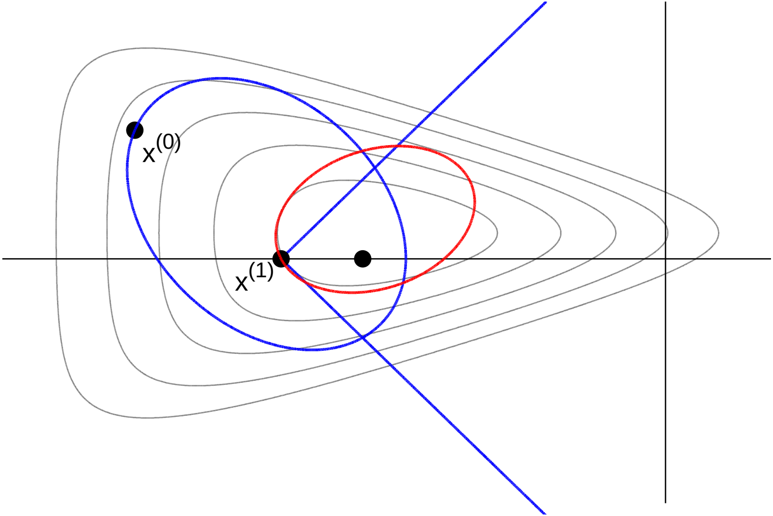
Optimization example



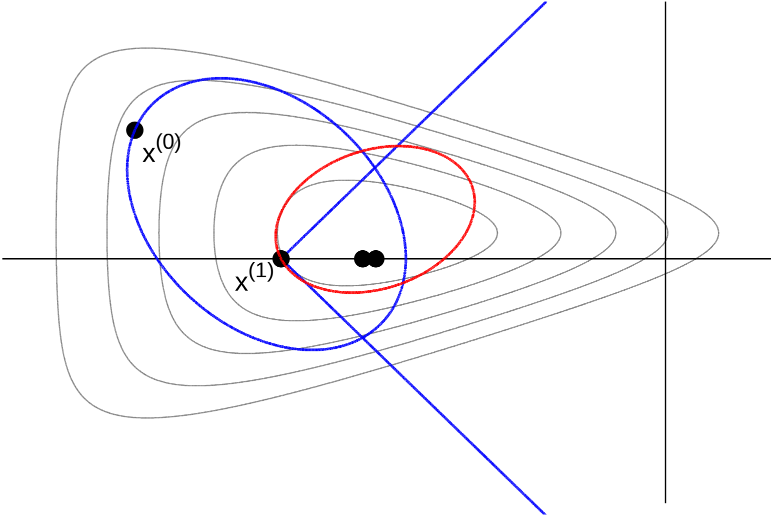
Optimization example



Optimization example

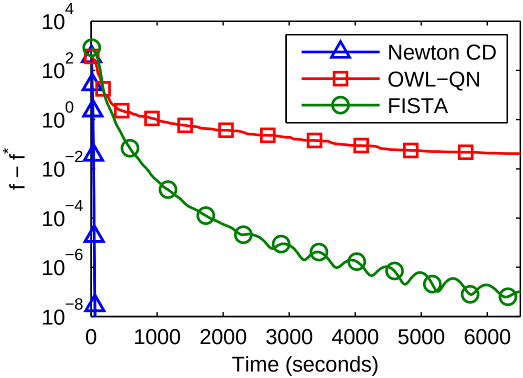


Optimization example



Optimization performance

Synthetic data with sparse underlying model, $n = 4000$, $p = 1000$



Converges to high numerical precision within 81 seconds while previous approaches require several hours

Theoretical results

- **Theorem**. Under proper assumptions and sample size

$$m = \Omega(d^4(\log p + \log n))$$

  where d is the max degree of the CRF, we have with high probability

  1. **Exact subset recovery**. The estimated parameters $\hat{\Lambda}, \hat{\Theta}$ have support that is a strict subset of the support of $\Lambda^\star, \Theta^\star$.

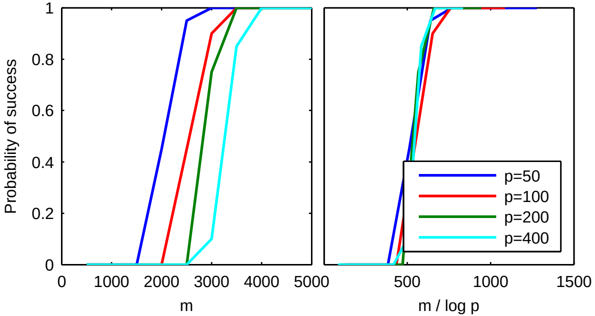
  2. ℓ_∞ **elementwise bound**.

$$\max(\|\hat{\Lambda} - \Lambda^\star\|_\infty, \|\hat{\Theta} - \Theta^\star\|_\infty) = O\left(\sqrt{\frac{\log p + \log n}{m}}\right)$$

- Based on the Primal-Dual Witness approach of Wainwright (2009) and Ravikumar et al. (2011)

Exact subset recovery

Chain CRF with bounded degree but growing p

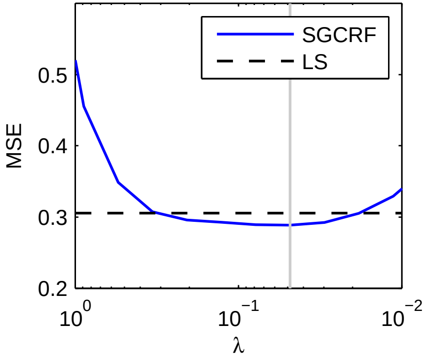


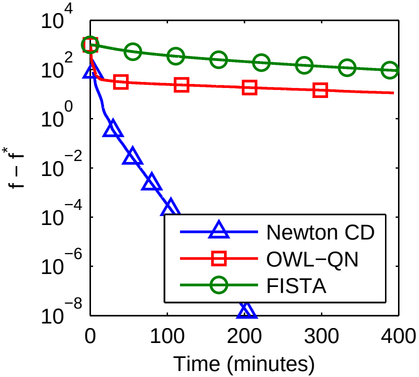
Rescaling the sample size demonstrates logarithmic dependence for exact subset recovery in accordance with theoretical results

Wind power forecasting

- Data from competition on Kaggle that ran in October 2012
- Outputs: wind power at 7 wind farms over 48 hours, $p = 336$
- Inputs: past 8 hours of wind power and 10 RBF features over wind forecasts, $n = 3417$
- Heavily optimized features for competition, resulting in a 5th place finish using ordinary least squares

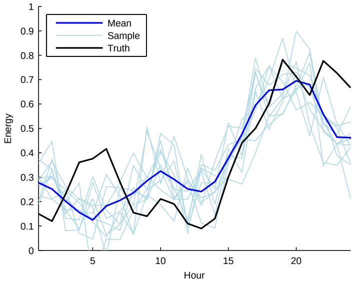
Wind power forecasting



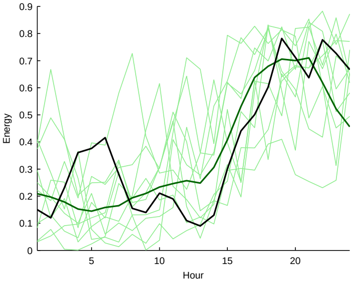


Improves on 5th place Kaggle entry by 5.5%

Wind power scenarios



Least squares

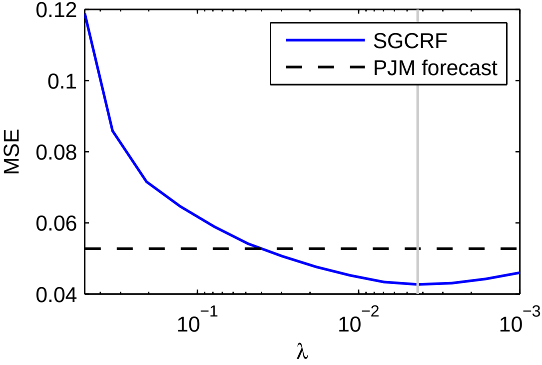


Sparse Gaussian CRF

Real advantage comes in accurately modeling the *distribution* over possible scenarios

Electrical demand forecasting

Predict energy demand in 15 zones over 24 hours, data from PJM, the electrical operator in Pennsylvania



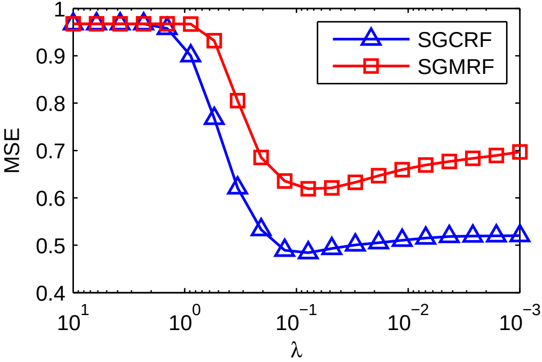
Improves on PJM's deployed system by 19%

Summary

- The sparse Gaussian CRF efficiently models dependencies of a *conditional* distribution
- We develop a second-order active set algorithm several orders of magntitude faster than previous approaches
- We provide theoretical analysis which characterizes statistical rates for graphs with bounded degree
- We achieve state-of-the art results in energy forecasting

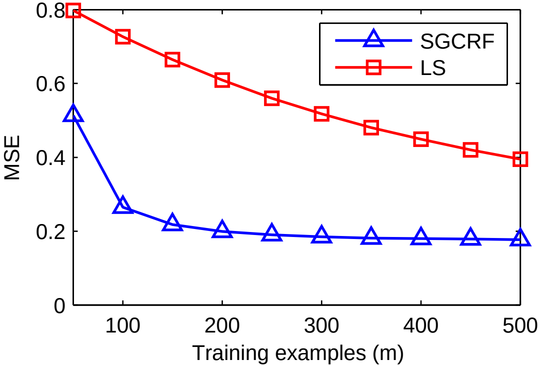
Comparison to MRF

Guassian distribution with sparse dependencies between y's and from y to x but not between x's



Does significantly better than the generative approach of modeling the full covariance of x, y

Sample size



The ℓ_1 penalty does much better than ℓ_2 regularized least-squares estimation when number of samples is small relative to features