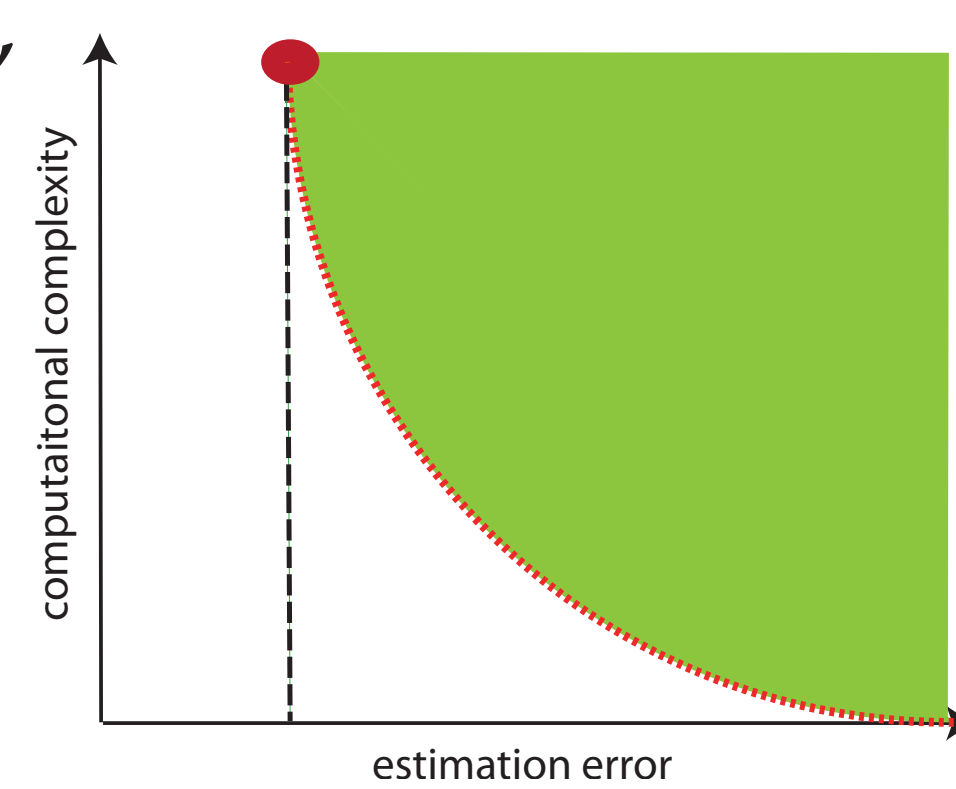


Introduction

- ▶ maximum likelihood estimators (mle) are a powerful technique for learning but are often limited in practice by intensive computation
- ▶ this work establishes a family of estimators that maximize a **stochastic** variation of the **composite likelihood** function
- ▶ prove the **consistency** of the estimators, provide formulas for their **asymptotic variance** and **computational complexity**
- ▶ experimental results for Boltzmann machines and conditional random fields (CRF) demonstrate the effectiveness of the estimators

Motivation

- ▶ mle possesses favorable qualities; consistent, i.e. $\hat{\theta}_n^{ml} \rightarrow \theta_0$ and has smallest possible asymptotic variance $(nI(\theta_0))^{-1}$ (Cramer-Rao)
- ▶ mle is often computationally intractable and motivates approx., e.g. pseudolikelihood
- ▶ such approx. cannot realize arbitrary computational/accuracy budgets (green)



Definition (m -pair)

An m -pair (A, B) is a pair of sets $A, B \subset \{1, \dots, n\}$ satisfying $A \neq \emptyset = A \cap B$. The likelihood object associated with an m -pair (A, B) and X is $S_{\theta}(A, B) = \log p_{\theta}(X_A | X_B)$ where $X_S \stackrel{\text{def}}{=} \{X_j : j \in S\}$. We similarly define likelihood objects with respect to a dataset $D = \{X^{(1)}, \dots, X^{(n)}\}$ as

$$S_{\theta}(n, A, B) = \sum_{i=1}^n \log p_{\theta}(X_A^{(i)} | X_B^{(i)}). \quad (1)$$

Definition (scl)

The stochastic composite loglikelihood (scl) associated with a finite sequence of m -pairs $(A_1, B_1), \dots, (A_k, B_k)$ is

$$scl_n(\theta; D) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \beta_j Z_{ij} \log p_{\theta}(X_{A_j}^{(i)} | X_{B_j}^{(i)}). \quad (2)$$

where $\beta_j > 0$ and $Z_{ij} \sim \text{Ber}(\lambda_j)$ are independent binary Bernoulli rv with parameters $\lambda_j \in [0, 1]$.

Proposition (Consistency)

Let $\lambda \in [0, 1]^k$ and $(A_1, B_1), \dots, (A_k, B_k)$ be a sequence of m -pairs for which $\{(A_j, B_j) : \forall j \text{ such that } \lambda_j > 0\}$ ensures identifiability. We also assume that $\Theta \subset \mathbb{R}^r$ is an open set and $p_{\theta}(x) > 0$ and is continuous and smooth in θ . Then there exists a strongly consistent sequence of scl maximizers, i.e. $\hat{\theta}_n^{msl} \rightarrow \theta_0$ as $n \rightarrow \infty$ with prob 1.

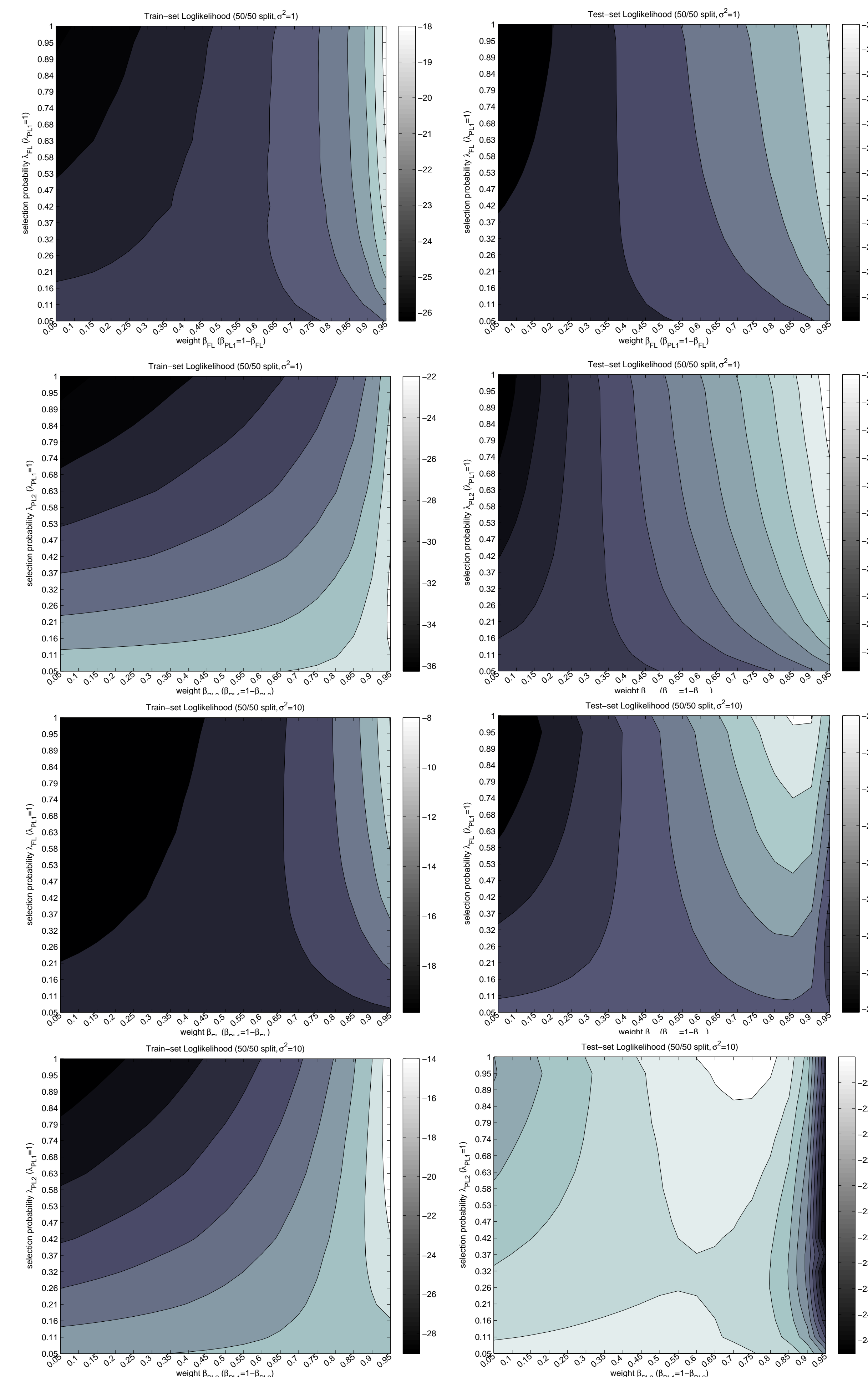
Proposition (Efficiency)

Making the assumptions of Proposition (consistency) as well as convexity of $\Theta \subset \mathbb{R}^r$ we have

$$\sqrt{n}(\hat{\theta}_n^{msl} - \theta_0) \rightsquigarrow N(0, Y \Sigma Y) \quad (3)$$

where $Y^{-1} = \sum_{j=1}^k \beta_j \lambda_j \text{Var}_{\theta_0}(V_j)$, $V_j = \nabla S_{\theta_0}(A_j, B_j)$, and $\Sigma = \text{Var}_{\theta_0}(\sum_{j=1}^k \beta_j \lambda_j V_j)$.

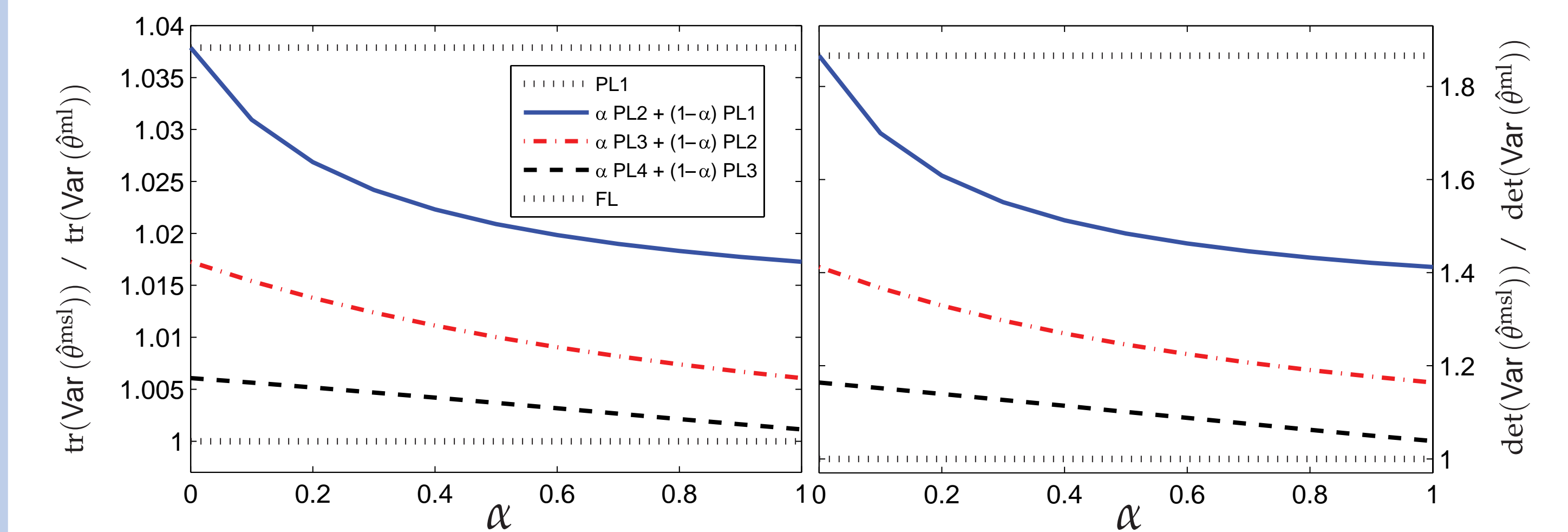
Conditional Random Field (real-world)



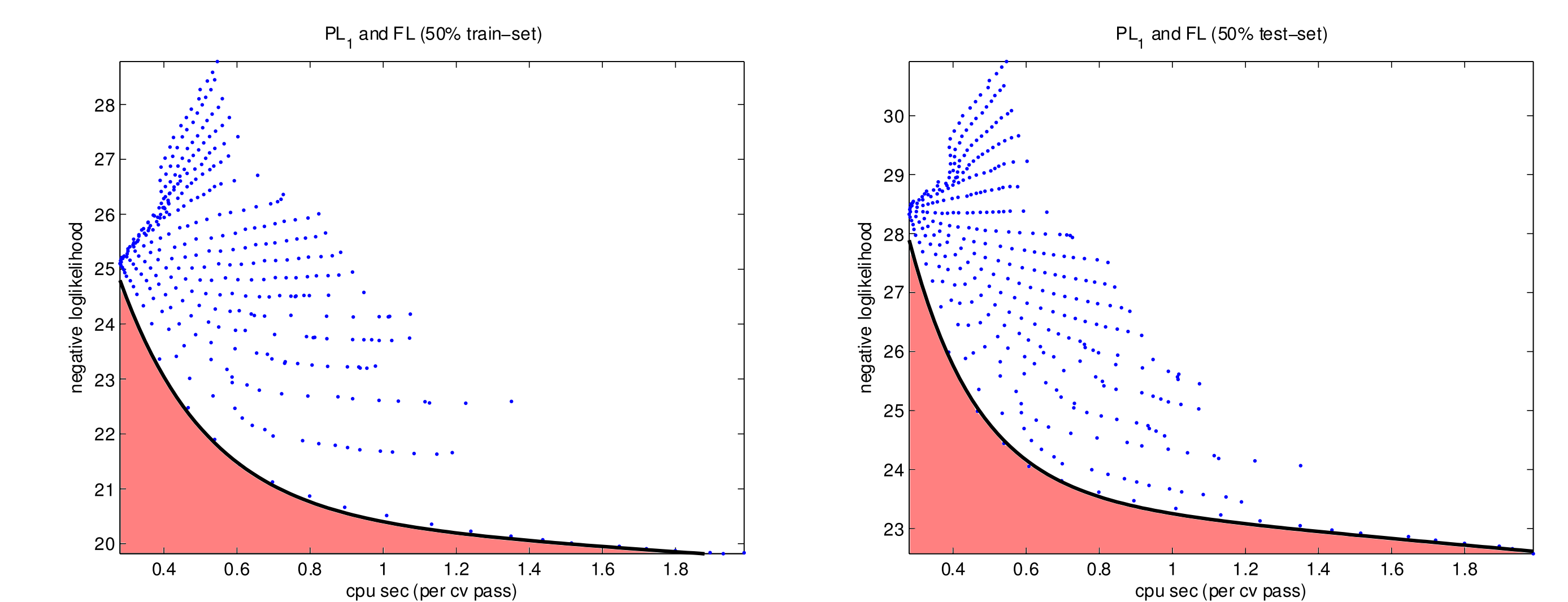
- ▶ train (left) and test (right) loglikelihood contours for maximum scl estimators for the CRF model
- ▶ L_2 regularization of $\sigma^2 = 1$ (rows 1,2) and $\sigma^2 = 10$ (rows 3,4)
- ▶ rows 1,3 are stochastic mixtures of full (FL) and pseudo (PL₁) likelihood components while rows 2,4 are pseudo (PL₁) and 2nd order pseudo (PL₂)

Boltzmann Machine (synthetic data)

- ▶ asymptotic variance matrix, as measured by trace (left) and determinant (right), as a function of the selection probabilities for different stochastic versions of the scl function.



CRF Accuracy/Complexity Tradeoff



- ▶ scatter plot representing complexity and negative loglikelihood (train on left, test on right) of scl functions for CRFs with regularization parameter $\sigma^2 = 1/2$
- ▶ points represent different stochastic combinations of full and pseudo likelihood components
- ▶ shaded region represents impossible accuracy/complexity demands

Conclusion

- ▶ proposed estimator family facilitates **computationally efficient** estimation in complex **graphical models**
- ▶ different parameterizations of the stochastic likelihood enables the resolution of the complexity-accuracy tradeoff in a domain and problem specific manner.
- ▶ framework is generally suited for **Markov random fields**, including conditional graphical models and is theoretically motivated
- ▶ in overfit models, stochastically mixing lower order components with higher order ones acts as a **regularizer** and results in a win-win situation of improving test-set accuracy and reducing computational complexity at the same time