Conditional Random Fields

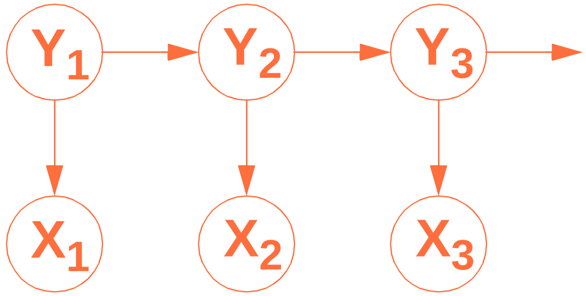
and

Decontamination

Motivation

HMM

- generative model
- specifies P(X,Y)
- can be used to compute P(X), P(Y|X), P(X|Y), etc.

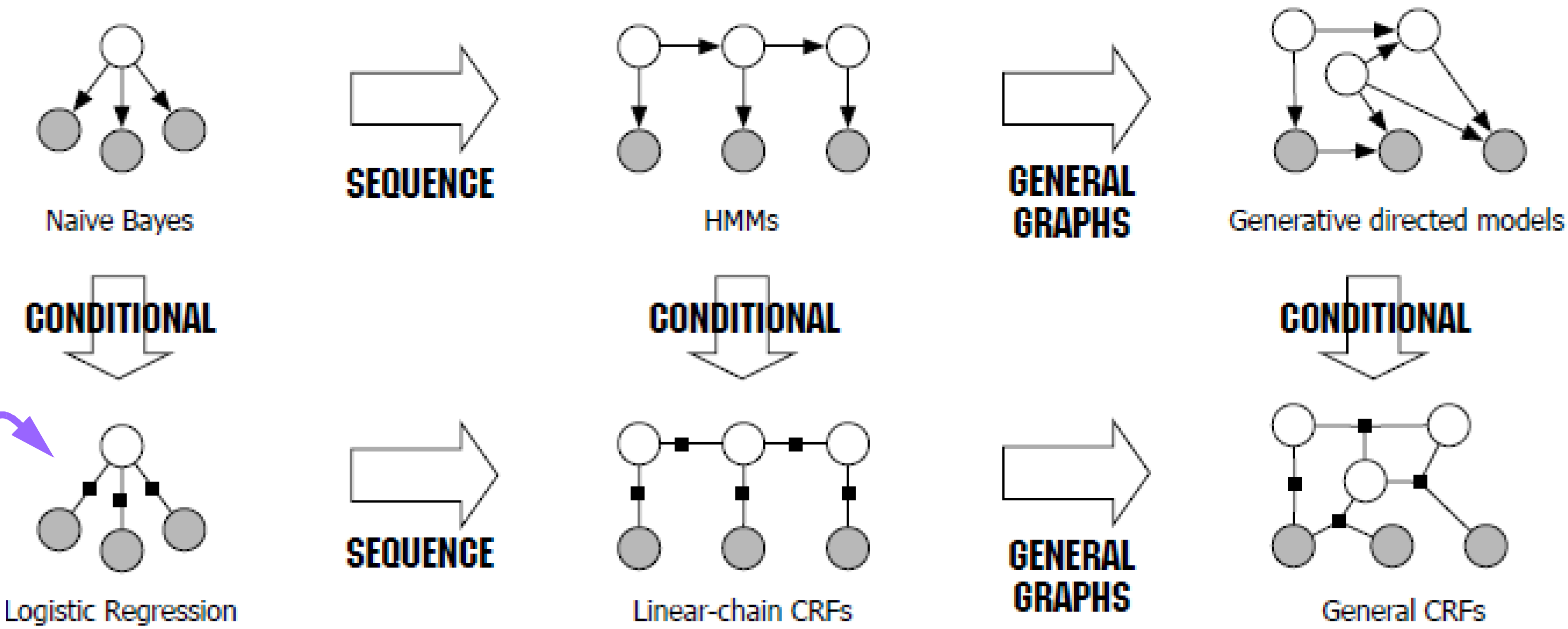


What do we typically do with HMM (e.g., speech recognition)?

What independence assumptions does it make?

- Given Y_3, what can we say about X_3 and X_2?
- Is this assumption sensible?

Relationships Among Models



factor graph

Lafferty, McCallum, & Pereira: Classic Paper Organization

1. existing approaches: HMM, MEMM

2. existing approaches have deficiencies

- assumption of independence of observations
- label bias problem

3. technique that overcomes the deficiencies: CRF

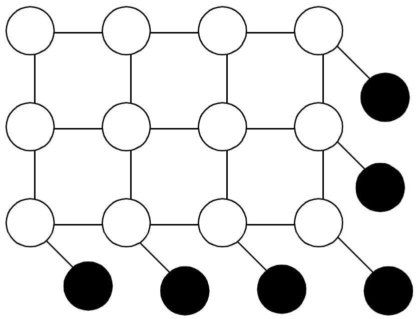
- general case
- special case (sequential structure)

4. algorithms for training CRF

5. simulations to show superiority of CRF over HMM, MEMM

- label bias
- mixed-order Markov model
- part-of-speech tagging

Markov Random Fields



E.g., Image segmentation

Undirected graphical model

Set of random variables, $\{Y_i\}$

Contextual constraints (spatial, temporal) connect neighbors

Neighborhood relations define cliques

- subsets of RVs in which every pair of distinct RVs are neighbors

Directed Vs. Undirected Graphical Model

Joint probability in *directed* graphical model:

$P(Y) = \prod_i P(Y_i \mid PA_i)$

i: index over nodes

Joint probability in undirected graphical model:

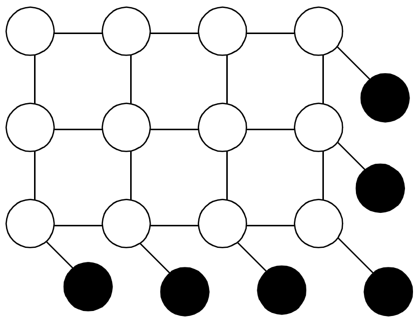
$P(Y) \sim \prod_c V_c(Y_c) = \exp(\sum_c \ln V_c(Y_c))$

c: index over cliques

Y_c: elements of Y in clique c (article uses $Y|_c$)

V_c: potential function that depends on configuration of clique

For discrete RVs, $V_c(Y_c) = \sum_{y_c} \underbrace{\lambda_{c,y_c}}_{\text{‘goodness’ of configuration}} \underbrace{f_{y_c}(Y_c)}_{\text{binary ‘feature’ (is } Y_c = y_c\text{?)}}$



Conditional Random Field

MRF specifies joint distribution on Y

For any probability distribution, you can condition it on some other variables X

CRF = MRF conditioned on X

MRF: $P(Y_i \mid Y_j \text{ for all } j \neq i) = P(Y_i \mid N_i)$ where N_i are the neighbors of i

CRF: $P(Y_i \mid X, Y_j \text{ for all } j \neq i) = P(Y_i \mid X, N_i)$ where N_i are the neighbors of i

$$P(Y \mid X) \sim \exp\left(\sum_{c,\, y_c} \lambda_{c,\, y_c} f_{y_c}(\{X, Y\}_c) \right)$$

CRF For Sequential Data

Framework

sequence of observations $X = \{X_i \text{ for } i = 1...n\}$

sequence of labels $Y = \{Y_i \text{ for } i = 1...n\}$

goal: infer Y given X

Applications

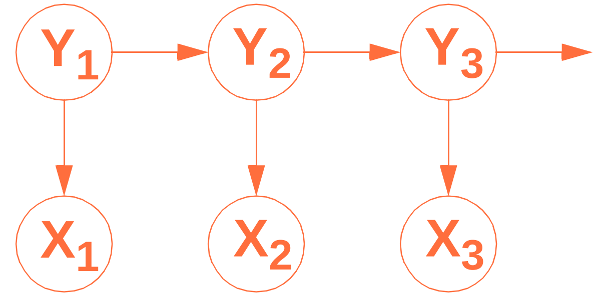
speech recognition

part of speech tagging

pretty much anything we use an HMM for, because it is typical to be given observation sequence X

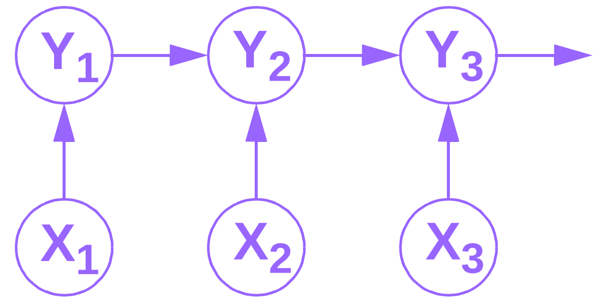
Relation To Other Sequential Models

HMM



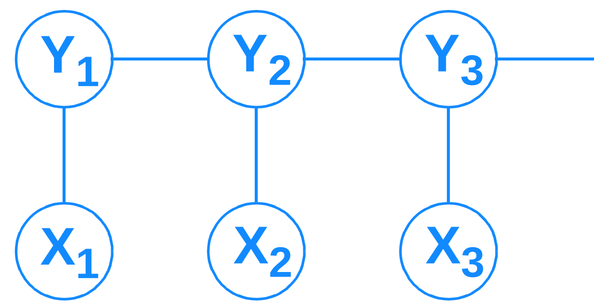
- generative model
- assumes cond. independence of X's
- does generality matter for recognition problems?

MEMM (Maximum Entropy Markov Model)



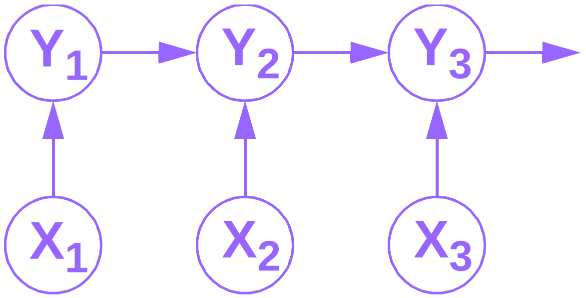
- specifies P(Y|X)
- does *not* require independence of X's
- many free parameters
- label bias problem

CRF



- specifies P(Y|X)
- does *not* require independence of X's
- fewer free parameters
- not subject to label bias problem

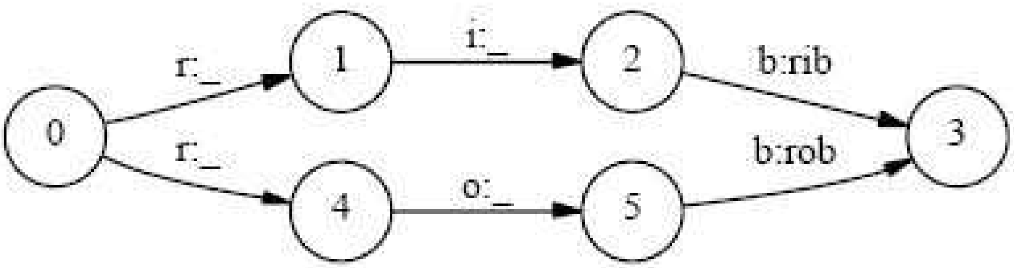
Label Bias Problem



Consider $P(Y_{i+1} \mid X_i, Y_i)$ in MEMM:

$\Sigma_j P(Y_{i+1}=j \mid X_i, Y_i) = 1$.

If only one possible value of Y_{i+1} is given Y_i, evidence is ignored.



Claim

Less robust to inaccurate modeling assumptions than CRF

Training Objective

HMM

Given observation sequence $\{X_1, X_2, ..., X_N\}$

Search for model parameters that maximize the likelihood of the observations.

$$L(\theta|X) = \prod_{i=1}^{N} P(X|\theta)$$

CRF

Given an observation sequence $\{(X_1,Y_1),(X_2,Y_2),...(X_N,Y_N)\}$

Search for model parameters that maximize the likelihood the conditional sequence

$$L(\theta|X, Y) = \prod_{i=1}^{N} P(Y|\theta, X)$$

Training Procedure

Algorithm exactly analogous to training procedure for HMM

1. Run forward-backward algorithm to obtain $P(Y_i|X,\theta)$

2. Adjust θ such that the inferred Y_i better match the training states

Simulation Studies

Label bias problem

CRF error 4.6%, MEMM error is 42%

how do they measure accuracy?

Synthetic data

from mixture of first- and second-order models

We generate data from a mixed-order HMM with state transition probabilities given by $p_\alpha(\mathbf{y}_i \mid \mathbf{y}_{i-1}, \mathbf{y}_{i-2}) = \alpha\, p_2(\mathbf{y}_i \mid \mathbf{y}_{i-1}, \mathbf{y}_{i-2}) + (1-\alpha)\, p_1(\mathbf{y}_i \mid \mathbf{y}_{i-1})$ and, similarly, emission probabilities given by $p_\alpha(\mathbf{x}_i \mid \mathbf{y}_i, \mathbf{x}_{i-1}) = \alpha\, p_2(\mathbf{x}_i \mid \mathbf{y}_i, \mathbf{x}_{i-1}) + (1-\alpha)\, p_1(\mathbf{x}_i \mid \mathbf{y}_i)$.

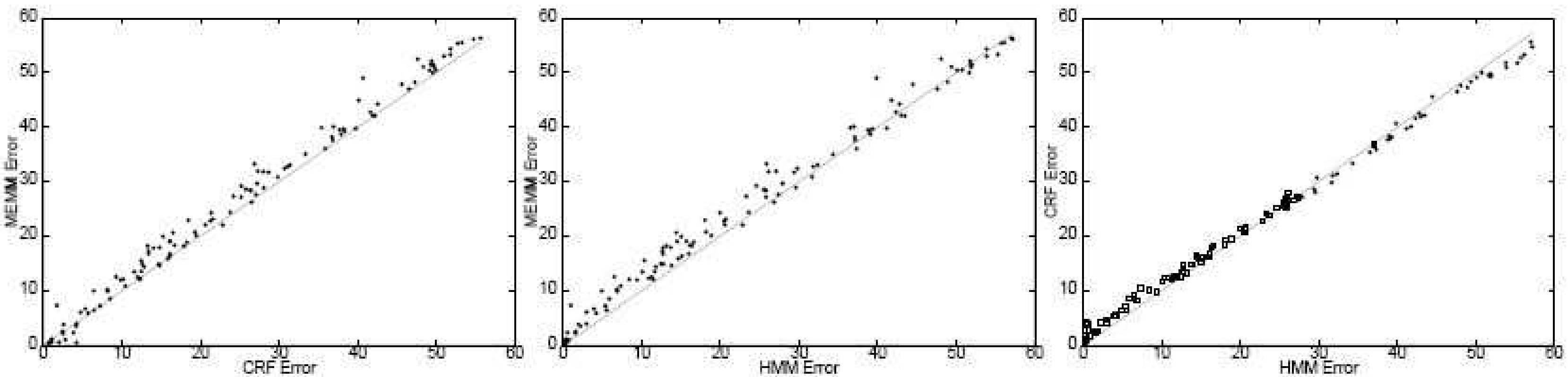


Figure 3. Plots of 2×2 error rates for HMMs, CRFs, and MEMMs on randomly generated synthetic data sets, as described in Section 5.2. As the data becomes "more second order," the error rates of the test models increase. As shown in the left plot, the CRF typically significantly outperforms the MEMM. The center plot shows that the HMM outperforms the MEMM. In the right plot, each open square represents a data set with $\alpha < \frac{1}{2}$, and a solid circle indicates a data set with $\alpha \geq \frac{1}{2}$. The plot shows that when the data is mostly second order ($\alpha \geq \frac{1}{2}$), the discriminatively trained CRF typically outperforms the HMM. These experiments are not designed to demonstrate the advantages of the additional representational power of CRFs and MEMMs relative to HMMs.

Part-of-speech tagging

label each word in English sentence with one of 45 syntactic tags

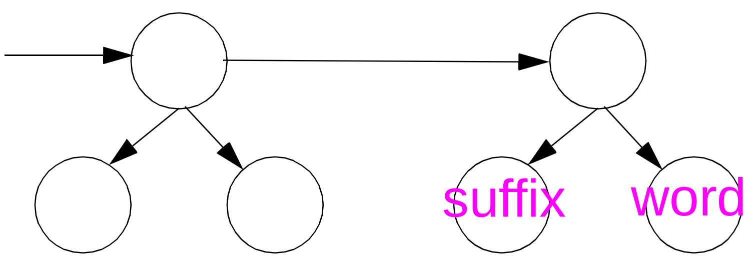
results for first-order HMM, MEMM, CRF in first 3 rows

model	*error*	*oov error*
HMM	5.69%	45.99%
MEMM	6.37%	54.61%
CRF	5.55%	48.05%
MEMM^{+}	4.81%	26.99%
CRF^{+}	4.27%	23.76%

$^{+}$Using spelling features

addition of orthographic features: begins with a number or upper case letter, contains a hyphen, whether it ends in specific suffix (-ing, -ogy, -ed, -s, -ly, etc.)

additional features rely on conditional *non*independence, vs. HMM



Decontaminating Human Judgments by Removing Sequential Dependencies

Michael C. Mozer
University of Colorado

Hal Pashler
UCSD

Matt Wilder, Rob Lindsey, Matt Jones
University of Colorado

Mike Jones
Indiana University

Sequential Effects In Judgment

Sequential Effects In Judgment

On a 1-10 scale, make a moral judgment about the following actions, with 1 indicating ‘not particularly bad or wrong’, and 10 indicating ‘extremely evil’:

(1) Stealing a towel from a hotel

(2) Keeping a dime you found on the ground

(3) Poisoning a barking dog

Sequential Effects In Judgment

On a 1-10 scale, make a moral judgment about the following actions, with 1 indicating ‘not particularly bad or wrong’, and 10 indicating ‘extremely evil’:

(1) Stealing a towel from a hotel

(2) Keeping a dime you found on the ground

(3) Poisoning a barking dog

Suppose instead the sequence had been:

(1’) Testifying falsely for pay

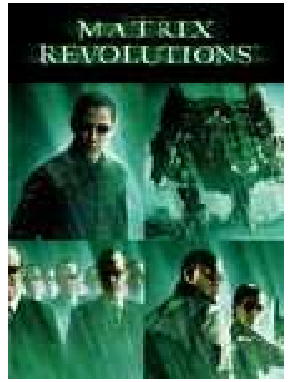
(2’) Using guns on striking workers

(3’) Poisoning a barking dog

Rating of action (3) is reliably higher than rating of action (3’) (Parducci, 1968)

Sequential Effects In Judgment

Rate these movies on a 1-5 scale



Netflix competition

- anchoring effects (early vs. late in rating session)
- slow drift

If ratings are contaminated by trial-to-trial sequence, can we *decontaminate* ratings to get scores that are more meaningfully related to an individual's internal sensation/impression/ evaluation?

Strategy

1. Collect data on a simple judgment task for which we have ground truth knowledge of the subjects' internal sensations

2. Use half of the subjects (*training subjects*) to build statistical/ probabilistic models of decontamination

3. Evaluate abilility of models to decontaminate on the other half of subjects (*test subjects*)

Sequential Effects:
Cognitive Models Vs. Decontamination Models

Cognitive model

Given past sequence of stimuli and responses, and current stimulus, predict current response (or response latency)

S(1), S(2), ... S(t), R(1), R(2), ...R(t-1) $\Rightarrow$ R(t)

Decontamination model

Given complete sequence of responses, predict complete sequence of sensations

R(1), R(2), ..., R(T) $\Rightarrow$ S(1), S(2), ..., S(T)

Gap Detection Task

On a 1-10 scale, judge how big the gap is between these two dots:

Gap Detection Task

On a 1-10 scale, judge how big the gap is between these two dots:

Gap Detection Task

On a 1-10 scale, judge how big the gap is between these two dots:

Gap Detection Task

On a 1-10 scale, judge how big the gap is between these two dots:

Absolute identification task (10 stimuli, 10 responses)

10 initial trials where subject is shown all 10 stimuli and is told the correct response

No further feedback

Like 1000's of similar studies in the literature, except without any feedback to make it more like Netflix rating task.

Experiments

Experiment 1

180 trials, 2 blocks of 90 trials

Within each block, all pairs of {gap(t-1), gap(t)} presented once, excluding repetitions

Within each 10 trials of block, all gaps presented once

gap = .08 K (K = 1, ..., 10)

Experiment 2

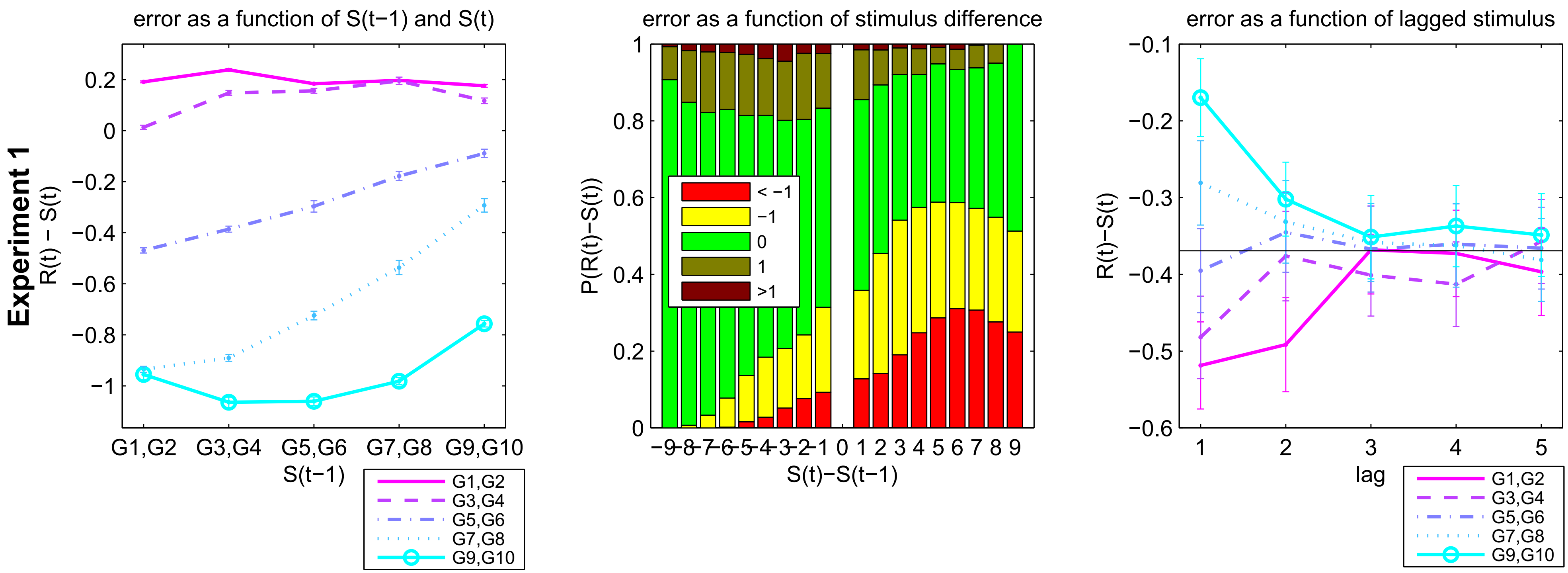
200 trials, 2 blocks of 100 trials

Within each block, all pairs of {gap(t-1), gap(t)} presented once, including repetitions

No subblock structure

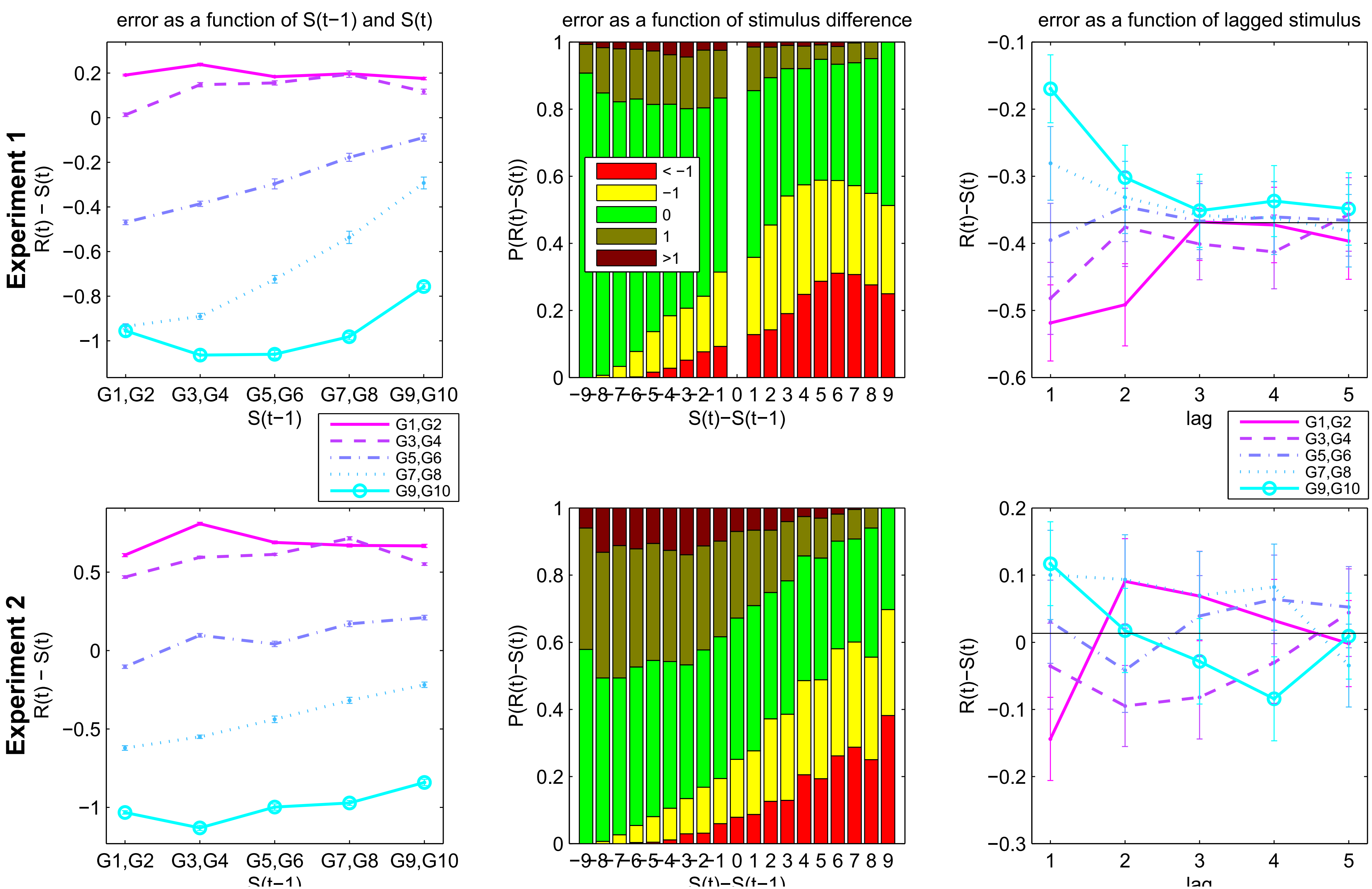
gap = .06 + .08 K (K = 1, ..., 10)

Results



Experiment 1
error as a function of S(t−1) and S(t)
R(t) − S(t)
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
G1,G2
G3,G4
G5,G6
G7,G8
G9,G10
S(t−1)
G1,G2
G3,G4
G5,G6
G7,G8
G9,G10
error as a function of stimulus difference
P(R(t)−S(t))
1
0.8
0.6
0.4
0.2
0
< −1
−1
0
1
>1
−9 −8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7 8 9
S(t)−S(t−1)
error as a function of lagged stimulus
R(t)−S(t)
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
1
2
3
4
5
lag
G1,G2
G3,G4
G5,G6
G7,G8
G9,G10

Results


error as a function of S(t−1) and S(t)
error as a function of stimulus difference
error as a function of lagged stimulus
Experiment 1
Experiment 2
R(t) − S(t)
P(R(t)−S(t))
R(t)−S(t)
S(t−1)
S(t)−S(t−1)
lag
G1,G2
G3,G4
G5,G6
G7,G8
G9,G10
< −1
−1
0
1
>1

Decontamination Models

1. Regression

$$\hat{S}(t) = \beta_0 + \beta_1 R(t) + \beta_2 R(t-1)$$

2. Look up table

$$\hat{S}(t) = \mathrm{Table}(R((t), R(t-1)))$$

3. Regression + look up table

$$\hat{S}(t) = \beta_0 + \beta_1 R(t) + \beta_2 R(t-1) + \mathrm{Table}(R((t), R(t-1)))$$

4. Conditional random field regression

$$\hat{S}(t) = \beta_0 + \beta_1 R(t) + \beta_2 R(t-1) + \beta_3 \hat{S}(t-1)$$

Inference via forward-backward algorithm

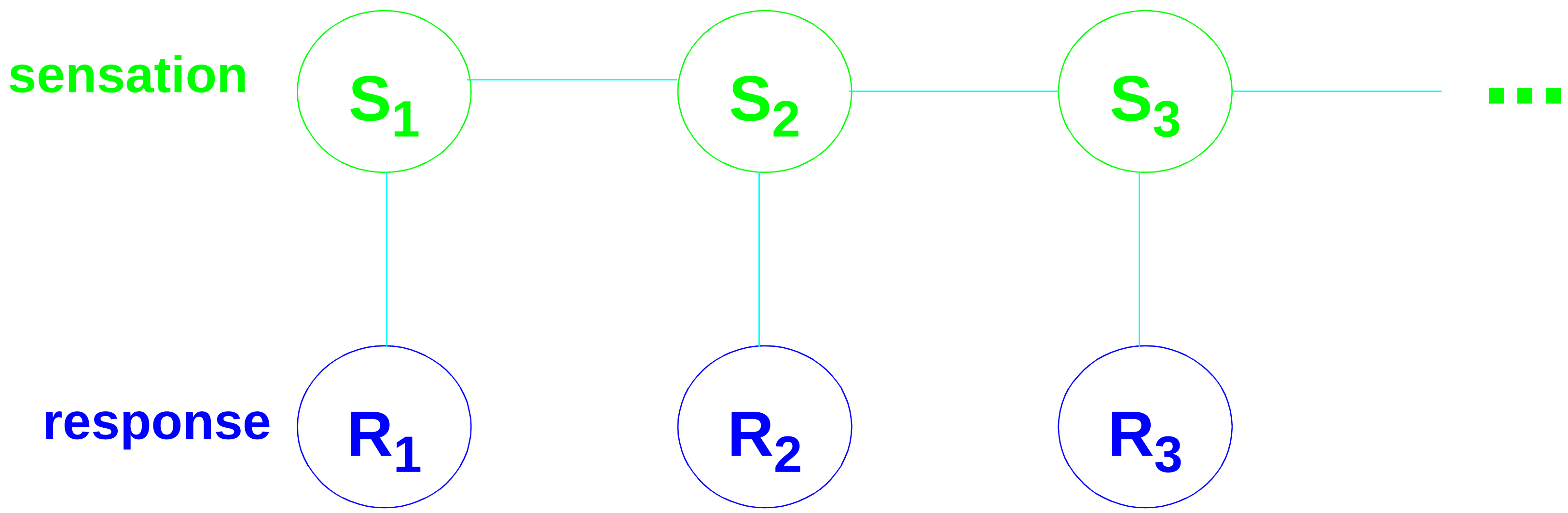
5. Conditional random field look up table

$$\hat{S}(t) = \mathrm{Table}(R(((t), R(t-1)), \hat{S}(t-1)))$$

6. Conditional random field regression + look up table

$$\hat{S}(t) = \beta_0 + \beta_1 Rr(t) + \beta_2 R(t-1) + \beta_3 \hat{S}(t-1) + \mathrm{Table}(R(((t), R(t-1)), \hat{S}(t-1)))$$

Conditional Random Fields



$$P(S_{1,T}|R_{1,T}) = \frac{1}{Z(R_{1,T})} \exp\left\{ \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(t, S_{t-1,t}, R_{1,T}) \right\}$$

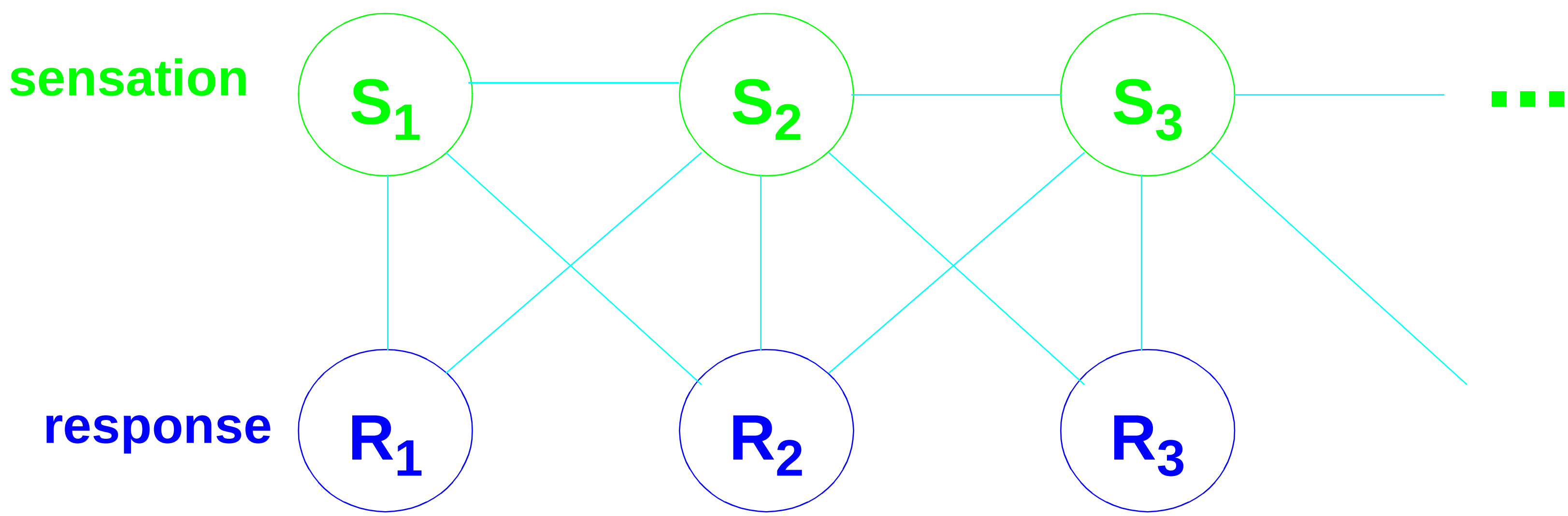
Conditional Random Fields

$$P(S_{1,T}|R_{1,T}) = \frac{1}{Z(R_{1,T})} \exp\left\{ \sum_{t=1}^{T} \boxed{\sum_{k=1}^{K} \lambda_k f_k(t, S_{t-1,t}, R_{1,T})} \right\}$$

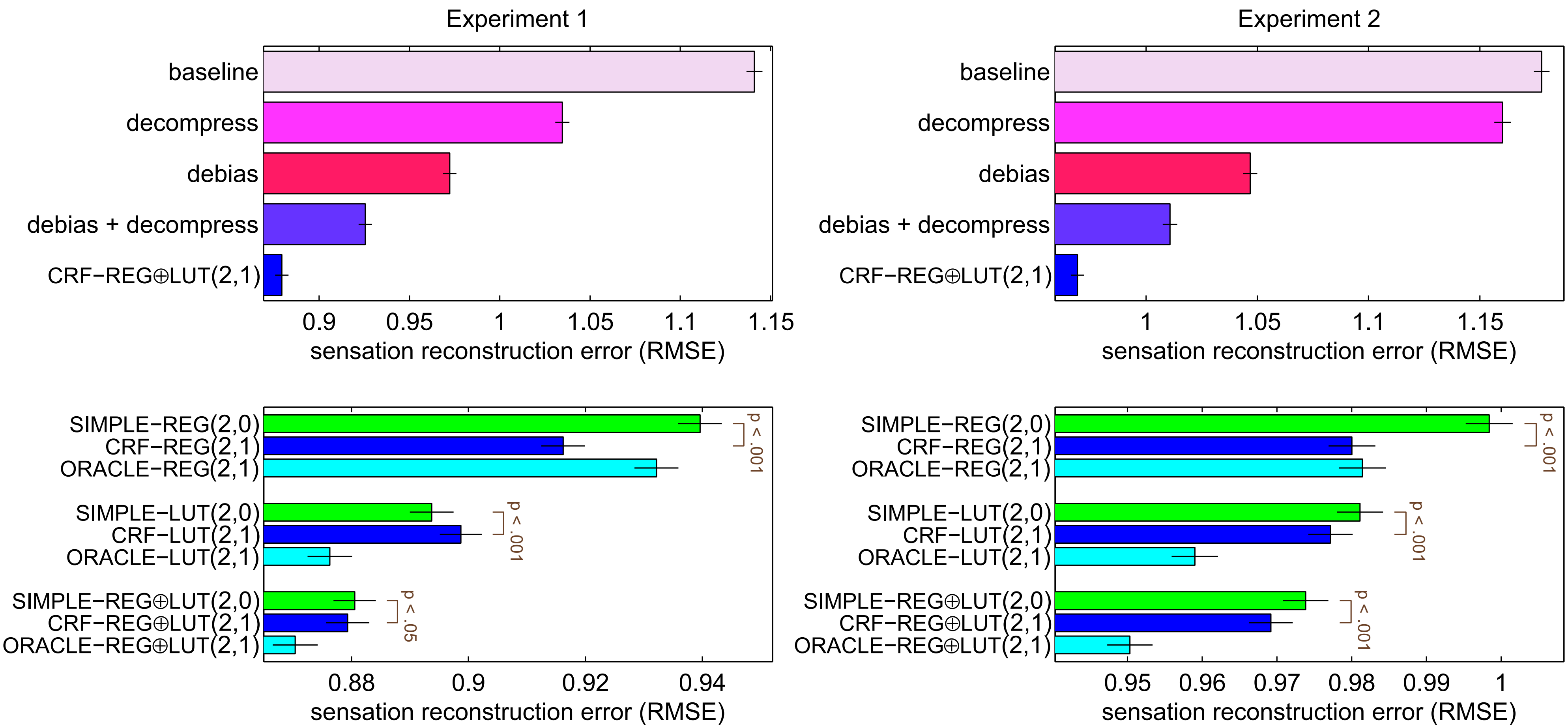
for regression: $\boxed{\Phi_t = -(\mathrm{REG}_t(m,n) - S_t)^2}$

$$\beta_0 + \beta_1 R(t) + \beta_2 R(t-1) + \beta_3 \hat{S}(t-1)$$

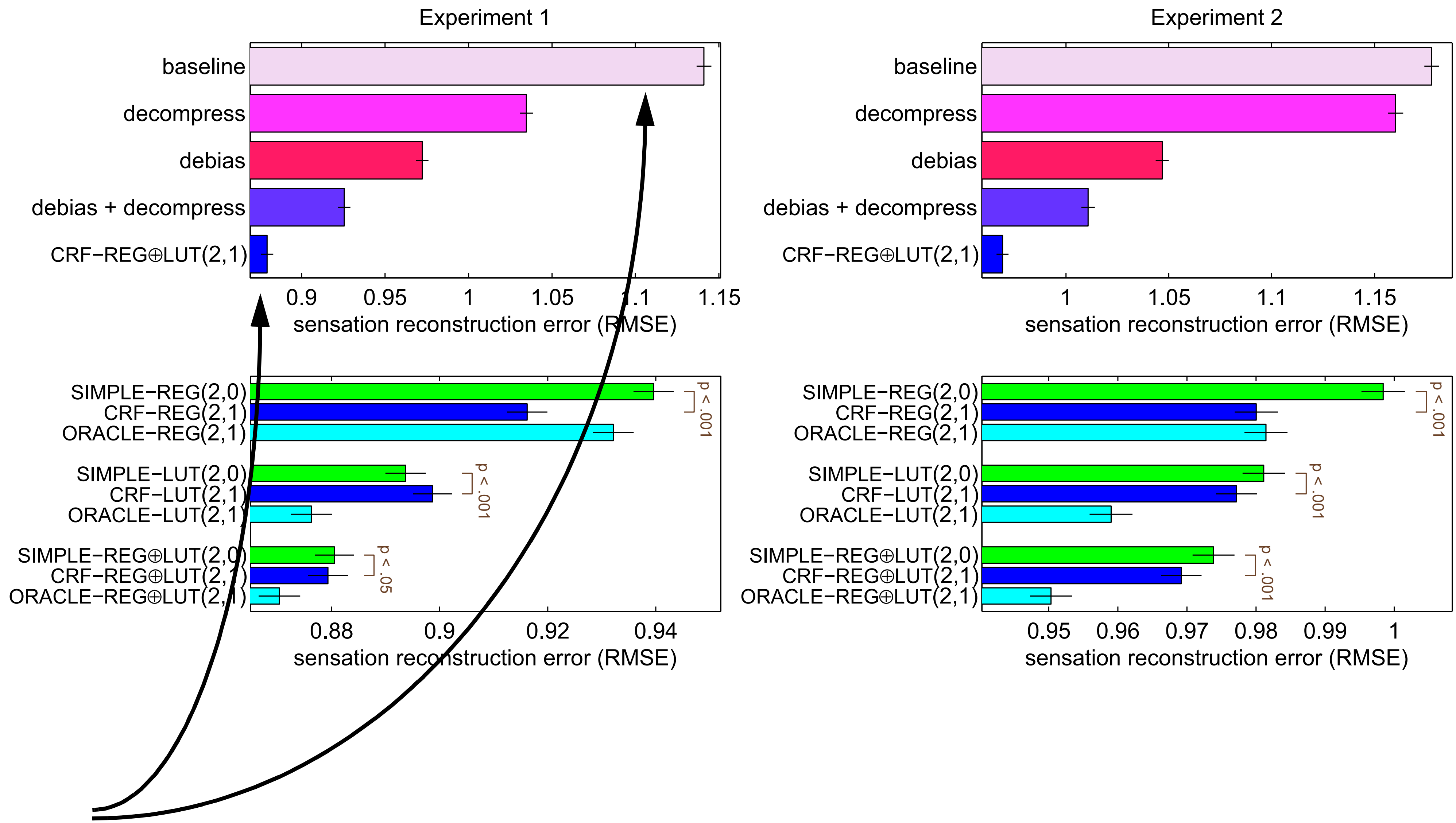
terms: S_t, R_tS_t, S_t^2, R_tS_{t-1}, $R_{t-1}S_t$, S_tS_{t-1}



Decontamination Results



Decontamination Results



Bottom line: 20% reduction in error over using subject's response vs. decontaminated estimate