# Learning From Measurements in Exponential Families

**Percy Liang** PLIANG@CS.BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA 94720, USA

**Michael I. Jordan** JORDAN@CS.BERKELEY.EDU
Computer Science Division and Department of Statistics, University of California, Berkeley, CA 94720, USA

**Dan Klein** KLEIN@CS.BERKELEY.EDU
Computer Science Division, University of California, Berkeley, CA 94720, USA

## Abstract

Given a model family and a set of unlabeled examples, one could either label specific examples or state general constraints—both provide information about the desired model. In general, what is the most cost-effective way to learn? To address this question, we introduce *measurements*, a general class of mechanisms for providing information about a target model. We present a Bayesian decision-theoretic framework, which allows us to both integrate diverse measurements and choose new measurements to make. We use a variational inference algorithm, which exploits exponential family duality. The merits of our approach are demonstrated on two sequence labeling tasks.

## 1. Introduction

Suppose we are faced with a prediction problem and a set of unlabeled examples. The traditional approach in machine learning is to label some of these examples and then fit a model to that labeled data. However, recent work has shown that specifying general constraints on model predictions can be more efficient for identifying the desired model (Chang et al., 2007; Mann & McCallum, 2008). In practice, one might want to use both labels and constraints, though previously these two sources have been handled in different ways. In this paper, we adopt a unified statistical view in which both labels and constraints are seen as ways of providing information about an unknown model.

Appearing in *Proceedings of the 26$^{th}$ International Conference on Machine Learning*, Montreal, Canada, 2009. 

| FEAT | FEAT | FEAT | FEAT | FEAT | ... |
|---|---|---|---|---|---|
| *View* | *of* | *Los* | *Gatos* | *Foothills* | ... |

| AVAIL | AVAIL | AVAIL | ... | SIZE | SIZE | SIZE | SIZE | ... |
|---|---|---|---|---|---|---|---|---|
| *Available* | *July* | *1* | ... | *2* | *bedroom* | *1* | *bath* | ... |

*Figure 1.* A sequence labeling task: Given a sequence of words from a Craigslist housing ad, label each word according to the type of information it provides: address, availability, contact, features, size, etc.

To this end, we introduce *measurements*, which subsume the notions of labels, partial labels, and general constraints on model predictions. Formally, a measurement is the expectation of a function (called a measurement feature) over the outputs of the unlabeled examples. A measurement provides a glimpse of the hidden outputs, thus providing partial information about the underlying model.

As a motivating application, consider the sequence labeling task shown in Figure 1. Given a Craigslist ad (a sequence of words) as input, the task is to output a label for each word indicating the semantic field to which it belongs (e.g., ADDRESS, SIZE, AVAILABILITY, etc.). Past research on this task has shown that in addition to obtaining labels of full sequences, it is particularly efficient to directly impose soft, cross-cutting constraints on the predictions of the model—for example, "the word *bedroom* is labeled as SIZE at least 90% of the time." Given both labels and constraints, how do we integrate them in a coherent manner? Additionally, how do we compare the value of information of various labelings and constraints?

To address these questions, we present a Bayesian decision-theoretic framework. Our setup follows the principles of Bayesian experimental design (see Chaloner and Verdinelli (1995) for an overview) but generalizes traditional designs in that we receive information not directly, via labeled data, but indirectly,

via measurements.

In order to scale up to large datasets, we present a variational approximation which exploits properties of Fenchel duality for exponential families. Our approximation is similar to the framework of Graça et al. (2008) for handling constraints on model predictions in a generative setting. Our variational objective can be optimized by solving a saddle point problem.

Empirically, we tested our method on a synthetic dataset and two natural language datasets (Craigslist ads and part-of-speech tagging), showing that we can integrate various types of measurements in a coherent way and also improve performance by actively selecting the measurements.

## 2. Measurements

Consider a prediction task, where $\mathcal{X}$ denotes the set of possible inputs and $\mathcal{Y}$ denotes the set of possible outputs. We start with a sequence of inputs $X = (X_1, \ldots, X_n)$, but unlike supervised learning, we do not observe the corresponding hidden outputs $Y = (Y_1, \ldots, Y_n)$. Instead, we propose making $k$ *measurements* on the data as follows:

$$\tau = \sum_{i=1}^{n} \sigma(X_i, Y_i) + W_\sigma, \tag{1}$$

where $\sigma(x, y) \in \mathbb{R}^k$ is a vector of *measurement features*, $\tau \in \mathbb{R}^k$ is a vector of (observed) *measurement values*, and $W_\sigma$ is some *measurement noise*.

The purpose of measurements is to provide a unified language for specifying partial information about $Y$. Traditional methods do deal with missing data, but the setting there is usually that of partial labels on individual examples. In contrast, we consider a more general space of mechanisms for partial supervision. Importantly, a measurement is an aggregate statistic that can span multiple examples.

In practice, measurement values $\tau$ can arise in two ways. The first is via *real measurements* obtained from the data; examples include labels of individual examples or aggregate values from a real experiment (e.g., pooling in genetics or census-taking). The second is via *pseudo-measurements*, which are set by hand to reflect prior knowledge, perhaps by "measuring" in a thought experiment. In binary classification, declaring that the fraction of positive examples is at least 95% is an example of a pseudo-measurement. The difference between pseudo- and real measurements is purely a conceptual one, as the two types are handled the same way inferentially.

We now give examples of measurements more formally:

**Fully-Labeled Example** To represent the output of some input $X_i \in \mathcal{X}$, let the components of $\sigma$ include $(x, y) \mapsto \mathbb{I}[x = X_i, y = b]$ for all $b \in \mathcal{Y}$.[1] Then the corresponding components of $\tau$ entirely determine $Y_i$. While these measurements are sums over all $n$ examples, $\tau$ can be computed by just inspecting $X_i$.

**Partially-Labeled Example** Suppose that we only observe $f(Y_i)$, a partial version of $Y_i$; for example, let $Y_i$ be a sequence and $f(Y_i)$ a subsequence. If $\sigma$ includes $(x, y) \mapsto \mathbb{I}[x = X_i, f(y) = b]$ for all $b \in f(\mathcal{Y})$, then $\tau$ reveals $Y_i$ up to $f(Y_i)$.

**Labeled Predicate** We can determine the outputs of all inputs $x$ for which $f(x) = 1$ by measuring $(x, y) \mapsto \mathbb{I}[f(x) = 1, y = b]$ for all $b \in \mathcal{Y}$. An example of a labeled predicate[2] in document classification occurs when $x$ is a document and $f(x) = 1$ if $x$ contains the word "market." Druck et al. (2008) showed that labeling these predicates can be more cost-effective than labeling full examples.

For sequence labeling tasks, we typically want to provide the frequency of some label $b$ over all positions where the input sequence is some $a$ (Mann & McCallum, 2008). For this, make measurements of the form $(x, y) \mapsto \sum_{i=1}^{\ell} \mathbb{I}[x_i = a, y_i = b]$, where $\ell$ is the length of the sequence.

In Quadrianto et al. (2008), examples are partitioned into sets, and all $(x, y) \mapsto \mathbb{I}[f(x) = a, y = b]$ are measured, where $f(x)$ is the set to which $x$ belongs.

**Label Proportions** Measuring $(x, y) \mapsto \mathbb{I}[y = b]$ for all $b \in \mathcal{Y}$ yields the proportions of each output label. This is the information used in expectation regularization (Mann & McCallum, 2007).

**Structured Label Constraints** Sometimes we have structural constraints on the outputs. In the Craigslist task, for instance, the output is a sequence $y = (y_1, \ldots, y_\ell)$. Domain knowledge tells us that each label either appears in a contiguous block or not at all. To capture this constraint, we use pseudo-measurement features $(x, y) \mapsto \sum_{i=1}^{\ell-1} \mathbb{I}[x = a, y_i = b, y_{i+1} = c]$ for each $a \in \mathcal{X}$ and labels $b \neq c$. We set the measurement values to $\tau = 0$ and their measure-

[1] The indicator function is $\mathbb{I}[a] = \begin{cases} 1 & \text{if } a = \text{true} \\ 0 & \text{otherwise.} \end{cases}$

[2] Druck et al. (2008) uses the term *feature* instead of *predicate*. We use *predicate* to denote an indicator function of the input $x$, reserving *feature* for functions on $(x, y)$.

ment noises to independent $-U[0,1]$. These quantities ensure that transitions into $b$, which mark the beginning of a new block for $b$, happen between 0 and 1 times (i.e., at most once). See Graça et al. (2008) for other types of constraints on structured outputs.

**Label Preferences** Suppose we don't know the exact proportions but strongly believe that $b^*$ is the most common label. This information can be encoded by the pseudo-measurement $(x,y) \mapsto \mathbb{I}[y=b^*] - \mathbb{I}[y=b]$ for $b \in \mathcal{Y}$ and setting $\tau = 0$ with noise $-U[0,n]$. These quantities ensure that $\sum_{i=1}^n \mathbb{I}[y=b^*] \geq \sum_{i=1}^n \mathbb{I}[y=b]$ for all $b \in \mathcal{Y}$. These preferences can also be adapted to operate conditioned on predicates.

It is often natural to obtain measurements of different types. We want to combine all the diverse measurements in a coherent way. This is important since there will naturally be varying amounts of redundancy across measurements. Furthermore, we would like a mechanism for determining which measurements to make next, accounting for both their costs and possible benefits. How to achieve these two goals in a principled way is the focus of the next section.

## 3. A Bayesian Framework

In this section, we present a Bayesian framework for measurements, which provides a unified way of both estimating model parameters given fixed measurements (Section 3.1) and optimally choosing new measurements (Section 3.2).

### 3.1. From Measurements to Model

Our goal is learn a predictor based on observed measurements. For the predictor, we use conditional exponential families, which include a broad class of prediction models, e.g., linear regression, logistic regression, and conditional random fields (Lafferty et al., 2001). A conditional exponential family distribution is defined as follows:

$$p_\theta(y \mid x) \stackrel{\text{def}}{=} \exp\{\langle \phi(x,y), \theta \rangle - A(\theta; x)\} \qquad (2)$$

for $x \in \mathcal{X}, y \in \mathcal{Y}$, where $\phi(x,y) \in \mathbb{R}^d$ is a vector of *model features*, $\theta \in \mathbb{R}^d$ is a vector of *model parameters*, and $A(\theta;x) = \log \int e^{\langle \phi(x,y),\theta \rangle} dy$ is the log-partition function.

Specifically, we would like to infer the model parameters $\theta$ from measurement values $\tau$ and inputs $X$. Recall that the outputs $Y$ are hidden. For guidance on how to perform this inference, we define the following Bayesian model (Figure 2(b)):

$$p(\theta, Y, \tau \mid X, \sigma) \stackrel{\text{def}}{=} p(\theta) \prod_{i=1}^n p_\theta(Y_i \mid X_i) p(\tau \mid X, Y, \sigma). \qquad (3)$$

For computational reasons, we assume the parameter prior and the noise distribution have log-concave densities: $\log p(\theta) = -h_\phi(\theta) + \text{constant}$ and $\log p(\tau \mid X, Y, \sigma) = -h_\sigma(\tau - \sigma^X(Y)) + \text{constant}$, where $g$ and $h$ are even convex functions, and $\sigma^X(Y) \stackrel{\text{def}}{=} \sum_{i=1}^n \sigma(X_i, Y_i)$. For example, we could use a Gaussian prior on $\theta$ ($h_\phi(\theta) = \frac{\lambda}{2}||\theta||^2$) and independent box noise ($h_\sigma(u) = \mathbf{W}[\forall j, |u_j| \leq \epsilon_j]$).[3]

Given (3), we can obtain the posterior $p(\theta \mid \tau, X, \sigma)$ by marginalization. It is conceptually useful to decompose this marginalization into two steps: We first combine pseudo-measurements $\tau_{\text{pseudo}} \subset \tau$ with a preliminary prior $p(\theta)$ to obtain a new prior $p(\theta \mid \tau_{\text{pseudo}}, X, \sigma)$. Then we combine this prior with real measurements to obtain the final posterior $p(\theta \mid \tau, X, \sigma)$. This situation is analogous to multinomial estimation with a conjugate Dirichlet prior: pseudocounts (concentration parameters) determine the prior, which combines with real counts to form the posterior.

We make one conceptual point regarding the relationship between measurement features and model features. While the two are the same type of mathematical object, they play different roles. Consider features $f_b(x,y) = \sum_{i=1}^\ell \mathbb{I}[\text{word } x_i \text{ ends in } \textit{-room}, y_i = b]$ for all labels $b$. As measurement features, $f$ would indicate that words ending in *-room* are likely to be labeled according to $\tau$. As model features, $f$ would indicate that words ending in *-room* are only labeled similarly. In this way, measurement features (along with $\tau$) provide direct information whereas model features provide indirect information. In general, measurement features should be finer-grained than model features, since finer features are easier to measure but coarser features generalize better.[4]

### 3.2. Active Measurement Selection

We now have a handle on how to learn from measurements, but how do we choose the optimal measurements $\sigma$ to make in the first place? To talk about optimality, we must define a utility function. For us, this involves predictive accuracy. First, define $r(y, \hat{y})$

[3] $\mathbf{W}[a] = \begin{cases} 0 & \text{if } a = \text{true} \\ \infty & \text{otherwise.} \end{cases}$

[4] A feature $f_1$ is *finer* than another feature $f_2$ if $f_2(x,y) = 0$ implies $f_1(x,y) = 0$.

to be the reward (e.g., label accuracy, or equivalently, negative Hamming loss) if the actual output is $y$ and we predict $\hat{y}$. If we use the Bayes-optimal predictor to make predictions on a new example $X'$ with true output $Y'$, the expected reward is as follows:

$$R(\sigma,\tau) \stackrel{\text{def}}{=} E_{p^*(X')} \max_{\hat{Y}'} E_{p(Y',\theta|X',\tau,X,\sigma)}[r(Y',\hat{Y}')]. \tag{4}$$

In short, $R(\sigma,\tau)$ measures our satisfaction with having made measurements $(\sigma,\tau)$. We also introduce $C(\sigma)$, the cost of measuring $\sigma$. Then the net (expected) utility is the difference:

$$U(\sigma,\tau) \stackrel{\text{def}}{=} R(\sigma,\tau) - C(\sigma). \tag{5}$$

In practice, we choose measurements in a sequential fashion. Suppose we have already made measurements $(\sigma_0,\tau_0)$ and want to choose the next $\sigma$ yielding the highest expected utility. However, since we do not know what measurement value $\tau$ we will obtain, we must integrate over $\tau$. Thus, the best subsequent measurement (feature) is given by the following:

$$\sigma^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\sigma\in\Sigma} U(\sigma), \tag{6}$$
$$U(\sigma) \stackrel{\text{def}}{=} E_{p(\tau|X,\sigma,\sigma_0,\tau_0)}[U((\sigma_0,\sigma),(\tau_0,\tau))],$$

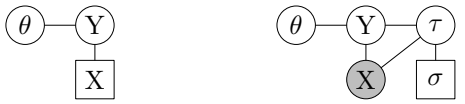
where $\Sigma$ is the set of candidate measurement features. Note that $\sigma^*$ is obtained via one-step lookahead, so it is only Bayes-optimal if $\sigma^*$ is the final measurement.

This completes the description of our measurement framework. Most of the computations above are intractable, so the remainder of this paper will focus on designing practical approximations.

We pause briefly to compare our framework with traditional experimental design (active learning). In both, there is an unknown parameter $\theta$ which governs $p_\theta(y \mid x)$. However, in traditional design, one chooses a set of inputs $X_1,\ldots,X_n$, whereupon the outputs $Y_1,\ldots,Y_n$ are revealed, and inference is then made on $\theta$. In our measurement framework, we choose measurement features $\sigma$, whereupon the measurement values $\tau$ are revealed, and inference is then made on $\theta$ through the latent variable $Y$, which must be integrated out. Figure 2 illustrates the distinction.

## 4. Approximation Methods

We now present methods for making the Bayesian principles described in the previous section practical. We first present an approximate inference algorithm for computing the posterior given fixed measurements (Section 4.1). We then present a method for actively choosing measurements (Section 4.2).




(a) Traditional design (b) Measurement design

*Figure 2.* In traditional experimental design (a), one selects $X$ and observes $Y$. In our measurement framework (b), one selects $\sigma$ and observes $\tau$.

### 4.1. Approximate Inference

Our approximate computation of the true posterior $p(Y,\theta \mid \tau,X,\sigma)$ proceeds in three steps. First, we apply a standard mean-field factorization (Section 4.1.1). Next, we relax the contribution of the measurements and apply Fenchel duality to obtain a workable objective function (Section 4.1.2). Finally, we present a strategy to optimize this function (Section 4.1.3).

#### 4.1.1. Mean-field factorization

Following standard variational principles, we turn (approximate) posterior computation into an optimization problem over a tractable set of distributions $\mathcal{Q}$:

$$\min_{q\in\mathcal{Q}} \mathrm{KL}\left(q(Y,\theta) \,||\, p(Y,\theta \mid \tau,X,\sigma)\right). \tag{7}$$

We use a mean-field approximation with a degenerate distribution over $\theta$:

$$\mathcal{Q} \stackrel{\text{def}}{=} \{q(Y,\theta) : q(Y,\theta) = q(Y)\delta_{\tilde{\theta}}(\theta)\}. \tag{8}$$

Now let us expand the original optimization problem (7) using (8) and (3):

$$\min_{q(Y),\theta} -H(q(Y)) + E_{q(Y)}[h_\sigma(\tau - \sigma^X(Y))] \tag{9}$$
$$-\sum_{i=1}^{n} E_{q(Y)} \log p_\theta(Y_i \mid X_i) + h_\phi(\theta).$$

#### 4.1.2. Relaxation and Fenchel duality

One problem is that the contribution of the measurements to the posterior (the second term of (9)) couples all the outputs $Y_1,\ldots,Y_n$. To progress towards tractability, we replace $E_{q(Y)}[h_\sigma(\tau - \sigma^X(Y))]$ in (9) with $h_\sigma(\tau - E_{q(Y)}[\sigma^X(Y)])$, which is a lower bound by Jensen's inequality. In doing so, we no longer guarantee a lower bound on the marginal likelihood.

However, this relaxation does let us rewrite (9) using Fenchel duality, which allows us to optimize over a vector $\beta \in \mathbb{R}^k$ rather than over an entire distribution $q(Y)$. Note that optimizing $q(Y)$ while holding $\theta$ fixed is exactly a maximum (cross-)entropy problem subject to approximate moment-matching constraints.

By Fenchel duality, the optimal $q(Y)$ belongs to an exponential family (cf. Dudík et al. (2007); Graça et al. (2008)):

$$q(Y) = \prod_{i=1}^{n} q_{\beta,\theta}(Y_i \mid X_i) \tag{10}$$

$$q_{\beta,\theta}(y \mid x) = \exp\{\langle \sigma(x,y), \beta\rangle + \langle \phi(x,y), \theta\rangle - B(\beta,\theta;x)\}, \tag{11}$$

where $B(\beta,\theta;x) = \log \int e^{\langle \sigma(x,y),\beta\rangle + \langle \phi(x,y),\theta\rangle} dy$ is the associated log-partition function for $q_{\beta,\theta}(y \mid x)$. See Appendix A for the derivation. We can now reformulate (9) as the following saddle point problem:

$$\min_{\theta\in\mathbb{R}^d} \max_{\beta\in\mathbb{R}^k} L(\beta,\theta), \tag{12}$$

$$L(\beta,\theta) = \langle \tau, \beta\rangle - \sum_{i=1}^{n} B(\beta,\theta;X_i) + \sum_{i=1}^{n} A(\theta;X_i) - h^*_\sigma(\beta) + h_\phi(\theta),$$

where $h^*_\sigma(\beta) = \sup_{u\in\mathbb{R}^k}\{\langle u,\beta\rangle - h_\sigma(u)\}$ is the Fenchel conjugate of $h_\sigma$. If we assume independent box noise, $h^*_\sigma(\beta) = \sum_j \epsilon_j|\beta_j|$.

Let $(\tilde\beta, \tilde\theta)$ be the solution to this saddle point problem. This pair specifies the approximate posterior $q(Y,\theta)$ via (10), (11), and (8).

#### 4.1.3. Optimization

We use a gradient-based approach to optimize $L(\beta,\theta)$ in (12). The gradients can be computed using standard moment-generating properties of log-partition functions:

$$\frac{\partial L(\beta,\theta)}{\partial \beta} = \tau - \sum_{i=1}^{n} E_{q_{\beta,\theta}(Y|X_i)}[\sigma(X_i,Y)] - \nabla h^*_\sigma(\beta), \tag{13}$$

$$\frac{\partial L(\beta,\theta)}{\partial \theta} = \sum_{i=1}^{n} E_{p_\theta(Y|X_i)}[\phi(X_i,Y)] - \sum_{i=1}^{n} E_{q_{\beta,\theta}(Y|X_i)}[\phi(X_i,Y)] + \nabla h_\phi(\theta). \tag{14}$$

Note that at $(\tilde\beta,\tilde\theta)$, both sets of moment-matching constraints (approximately) hold: (1) the measurement feature expectations under $q_{\beta,\theta}$ are close to the observed values $\tau$, indicating that we have represented the measurements faithfully; and (2) the model feature expectations under $q_{\beta,\theta}$ are close to those of $p_\theta$, indicating that we have learned a good model.

Because $n$ is typically large, we use stochastic approximations to the gradient. In particular, we take alternating stochastic gradient steps in $\beta$ and $\theta$. At the end, we return an average of the parameter values obtained along the way to provide stability.

The alternating quality of our algorithm is similar in spirit to Chang et al. (2007). However, they maintain a list of candidate $Y$s instead of a distribution $q(Y)$. They also use a penalty for violating constraints (the analog of our $\beta$), which must be manually set. We only require specifying the form of the measurement noise, which is more natural; from this, $\beta$ is learned automatically.

Note that the only computations we need are expected feature vectors, which are standard quantities needed in any case for gradient-based optimization procedures. In contrast, the use of Generalized Expectation Criteria requires computing the covariance between $\sigma$ and $\phi$, which is more complex and expensive for graphical models (Mann & McCallum, 2008).

#### 4.1.4. Intuitions

For simplicity, assume zero measurement noise and a flat improper prior on $\theta$. Let $\mathcal{P} = \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$ denote our model family. Let $\mathcal{Q}$ (different from before) be the set of all distributions which are consistent with our measurements $(\sigma,\tau)$. Our variational approximation can be interpreted as finding a $q_{\beta,\theta}(y \mid x) \in \mathcal{Q}$ and a $p_\theta(y \mid x) \in \mathcal{P}$ such that $\mathrm{KL}\,(q \,||\, p)$ is minimized.[5] Intuitively, all external information is fed into $\mathcal{Q}$, a staging area, which ensures we work with coherent distributions. This information is then transferred to our model family $\mathcal{P}$, which allows us to generalize beyond our observations.

When the measurement features are the same as the model features ($\sigma \equiv \phi$), the problem reduces to standard supervised learning by maximum entropy duality. In particular, $\mathcal{Q}\cap\mathcal{P}$ contains the unique solution. Another way to obtain supervised learning is to measure $(x,y) \mapsto \mathbb{I}[x = a, y = b]$ for all $a \in \mathcal{X}, b \in \mathcal{Y}$ (cf. Section 2). Then $\mathcal{Q}$ is a single point which typically lies outside $\mathcal{P}$.

Druck et al. (2008) incorporate measurements using Generalized Expectation Criteria, an objective function that penalizes some notion of distance (e.g., KL-divergence) between $E_{p_\theta(Y|X)}[\sigma(X,Y)]$ and the measurement values $\tau$. Even when $\sigma \equiv \phi$, their objective function does not reduce to supervised learning and

[5] In the language of information geometry, optimizing $q$ with $p$ fixed is an I-projection; optimizing $p$ with $q$ fixed is an M-projection.

thus does not resolve redundant measurements in a coherent way.

### 4.2. Approximate Active Measurement Selection

In traditional design, if $p_\theta(y \mid x)$ is a linear regression model and $p(\theta)$ is Gaussian, then $U(\sigma)$ has a closed-form expression. If a non-conjugate $p(\theta)$ is employed, for example, a sparsity prior for compressed sensing, one must resort to approximations such as expectation propagation (Seeger & Nickisch, 2008).

In our measurement setting, inference is further complicated by marginalization over $Y$, so let us apply our posterior approximations to measurement selection (Section 3.2). First, consider the expected reward (4). Since we do not have access to the true test distribution $p^*(x)$, we use a heldout set of unlabeled examples $\tilde{X}_1, \ldots, \tilde{X}_m$. Define $\tilde{p}(x) = \frac{1}{m}\sum_{i=1}^m \delta_{\tilde{X}_i}(x)$ to be the corresponding empirical distribution.

Second, we replace the true posterior $p(\theta \mid \tau, X, \sigma)$ with a point estimate $\tilde{\theta}(\sigma, \tau)$, thereby obtaining an approximate utility:

$$\tilde{U}(\sigma, \tau) \stackrel{\text{def}}{=} E_{\tilde{p}(X')} \max_{\hat{Y}'} E_{p_{\tilde{\theta}}(Y' \mid X')}[r(Y', \hat{Y}')] - C(\sigma). \tag{15}$$

The final step is to marginalize out $\tau$. Suppose we have already made measurements $(\sigma_0, \tau_0)$. The true posterior $p(Y, \theta \mid X, \sigma_0, \tau_0)$ is currently approximated by $q_0(Y, \theta)$ (represented by $(\tilde{\beta}_0, \tilde{\theta}_0)$). Using this approximation leads to $q_0(\tau) \stackrel{\text{def}}{=} E_{q_0(Y)}[p(\tau \mid X, Y, \sigma)]$ as a substitute for $p(\tau \mid X, \sigma, \sigma_0, \tau_0)$.

Though we can compute $\tilde{U}((\sigma_0, \sigma), (\tau_0, \tau))$ for a fixed $\tau$, we cannot integrate $\tilde{U}$ over $\tau$. Thus, we use a Monte Carlo approximation: Draw $t$ samples from $q_0(\tau)$ by first drawing $Y$ from $q_0(Y)$ and then sampling $\tau$ according to (1). Let $\tilde{q}_0(\tau)$ be the empirical distribution formed from these samples. Now $U(\sigma)$ from (6) can be then approximated with the following:

$$\tilde{U}(\sigma) = E_{\tilde{q}_0(\tau)}[\tilde{U}((\sigma_0, \sigma), (\tau_0, \tau))]. \tag{16}$$

The pseudocode for our algorithm is given in Figure 3. This procedure is similar in spirit to the active learning algorithm proposed by Roy and McCallum (2001), where examples were chosen iteratively to minimize expected loss on heldout data under a Naïve Bayes model.

One potential weakness with our approach is that we do not maintain any uncertainty in $\theta$ in the variational approximation. If we were doing parameter estimation, this approximation would be entirely useless since what drives experimental design in that case is the reduction of uncertainty in $\theta$. However, our utility function is predictive accuracy. Intuitively, what drives our method is reduction of uncertainty in predictions based on $\tilde{\theta}$. For this, the magnitude of $\tilde{\theta}$ does provide some guidance.

Algorithm for Active Measurement Selection

$\sigma_0 \leftarrow \emptyset \quad \tau_0 \leftarrow \emptyset \quad \tilde{\beta}_0 \leftarrow \emptyset \quad \tilde{\theta}_0 \leftarrow 0$
while more measurements are desired:
  for each candidate measurement feature $\sigma \in \Sigma$:
    draw $t$ samples from $q_0(\tau)$ specified by $(\tilde{\beta}_0, \tilde{\theta}_0)$
    for each sampled measurement value $\tau$:
      $(\tilde{\beta}, \tilde{\theta}) \leftarrow \textsc{ApproxInference}((\sigma_0, \sigma), (\tau_0, \tau))$
      $u_{\sigma,\tau} \leftarrow \tilde{E}_{\tilde{p}(X')} \max_{\hat{Y}'} E_{p_{\tilde{\theta}}(Y' \mid X')}[r(Y', \hat{Y}')] - C(\sigma)$
    $u_\sigma \leftarrow \frac{1}{t}\sum_\tau u_{\sigma,\tau}$
  $\sigma^* \leftarrow \text{argmax}_\sigma\, u_\sigma$
  obtain measurement value $\tau^* = \sigma^X(\sigma^*) + W_{\sigma^*}$
  $\sigma_0 \leftarrow (\sigma_0, \sigma^*) \quad \tau_0 \leftarrow (\tau_0, \tau^*)$
  $(\tilde{\beta}_0, \tilde{\theta}_0) \leftarrow \textsc{ApproxInference}(\sigma_0, \tau_0)$
Output $\tilde{\theta}_0$

*Figure 3.* Pseudocode for choosing measurements in a sequential manner based on our variational approximation.

## 5. Experiments

We now present empirical results. In Section 5.1, we show how the measurement framework can effectively integrate both labeled data and labeled predicates. In Section 5.2, we actively choose the measurements.

### 5.1. Learning from Measurements

For the Craigslist task introduced in Section 1, we use a linear-chain conditional random field (CRF), which is a conditional exponential family where the input $x = (x_1, \ldots, x_\ell)$ is a sequence of words, the output $y = (y_1, \ldots, y_\ell)$ is a sequence of labels, and the model features are $\phi(x, y) = \sum_{i=1}^\ell \phi^1(y_i, x, i) + \sum_{i=1}^{\ell-1} \phi^2(y_i, y_{i+1})$. The components of the node features $\phi^1(y_i, x, i)$ are indicator functions of the form $\mathbb{I}[y_i = a, s(x_i) = b]$, where $a$ ranges over the 11 possible labels, and $s(\cdot)$ is either the identity function or a function mapping each word to one of 100 clusters. To create these clusters, we ran the Brown word clustering algorithm.[6]

We started $n = 1000$ unlabeled examples and con-

[6]In order to capture topical similarity, three-word sequences $(x_{i-d}, x_i, x_{i+d})$ were created for each sequence, position $i = 1, \ldots, \ell$, and offset $d = 1, 2, 3$. The word clusters obtained from these three-word sequences essentially capture the same type of structure in the data as the SVD features used by Haghighi and Klein (2006) and Mann and McCallum (2008).

*Table 1.* CCR07 (Chang et al., 2007) and MM08 (Mann & McCallum, 2008) outperform our method when there are few examples, but we achieve the best overall number with 100 examples.

| # labeled examples | 10 | 25 | 100 |
|---|---|---|---|
| CCR07 | **74.7** | **78.5** | 81.7 |
| MM08 | 74.6 | 77.2 | 80.5 |
| no labeled predicates | 67.7 | 75.6 | 81.5 |
| + 33 labeled predicates | 71.4 | 76.5 | **82.5** |

sidered two types of measurements: fully-labeled examples and labeled predicates where we provide the frequency of the most common label for a word type.[7] The labeled examples were chosen at random and we chose three "prototypes" for each of the 11 labels based on the 100 available labeled training examples (see Haghighi and Klein (2006) for details).

We optimized $L(\beta, \theta)$ for 50 iterations ($50n$ stochastic steps). Table 1 shows that the performance of our method (100 examples) improves as we add more labeled examples and predicates. To the best of our knowledge, our 82.5% is the best result published so far on this task. More interestingly, compared to past work, we get larger gains as we label more examples, which suggests that our measurement framework is integrating the diverse, increasing information more effectively.

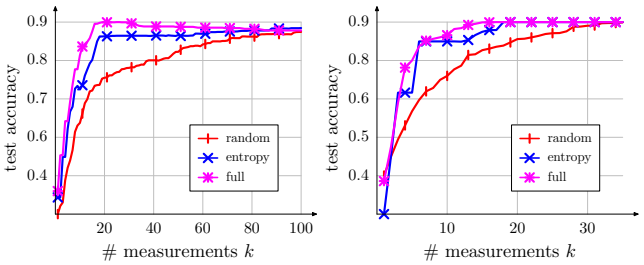
### 5.2. Active Measurement Selection

#### 5.2.1. Synthetic dataset

Consider the following multiclass classification problem: the output space is $\mathcal{Y} = \{1, \ldots, 4\}$, and the input space is $\mathcal{X} = \cup_{y \in \mathcal{Y}} \{(x_1, x_2) : i \in \{1, \ldots, 5\}, x_1 = (y, i), x_2 \in \{y\} \times \{1, \ldots, 2^{i-1}\}\}$. Inputs are generated uniformly from $\mathcal{X}$ and each input $x$ is assigned a label $y$ which is extracted from $x$ with probability 0.9 and uniformly from $\mathcal{Y}$ with probability 0.1. We consider two types of measurements: fully-labeled examples and labeled $x_2$-predicates. For simplicity, assume all measurements have the same cost.

We started with $n = 100$ unlabeled examples and no measurements. Following Figure 3, for each candidate measurement feature, we drew $t = 3$ samples of $\tau$. Given each hypothetical measurement, we ran approximate inference for 10 iterations, warm starting from the previous parameter setting $(\tilde{\beta}, \tilde{\theta})$. Our utility $\tilde{U}$ was computed on a heldout set of 500 examples. Then the best measurement feature was added, followed by

[7]The frequency was measured on the 100 available labeled examples and extrapolated to the rest. We assumed zero measurement noise.




(a) Labeling examples (b) Labeling word types

*Figure 4.* Comparison of three methods on the synthetic dataset: iteratively choosing the next measurement at random, based on entropy, or by running our full algorithm.

10 more iterations of approximate inference. Finally, we evaluated test accuracy on 1000 fresh examples.

We compared the full algorithm we just described with two alternatives: (1) choosing the next measurement at random and (2) choosing the example or predicate with the highest entropy.[8] Figure 4 shows the results, averaged over 10 trials. We see that both the entropy-based heuristic and the full algorithm provide substantial gains over random, and moreover, the full algorithm provides a slight edge over entropy. One property that entropy fails to capture is the propagation effect: Two measurements might have the same entropy, but they could have different degrees of impact on other examples through re-estimating the model. However, the full algorithm does come with a significant computational cost, so for the experiments in the next section we used entropy.

#### 5.2.2. Part-of-speech tagging

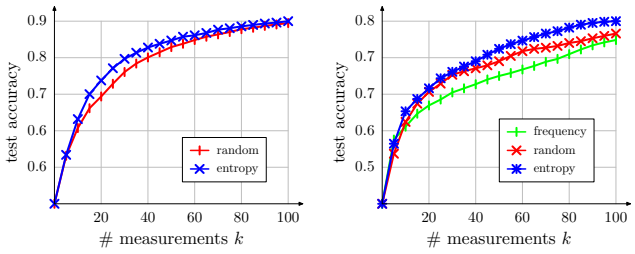
Now we turn to part-of-speech tagging.[9] Using standard capitalization, suffix, word form, and word cluster features applied on the previous, current, and next words, we seek to predict the tag of the current word. We considered two types of measurements: (1) tagging a whole sentence and (2) providing the frequency of the most common tag for a word type, where the word type is one of the 100 most frequent.

We started with 1000 unlabeled training examples and labeled 10 examples at random. Then we went through candidate measurements, evaluating them using entropy,[10] and adding the best five each round, after which a single iteration of approximate inference was

[8]The entropy of a predicate is the sum over the label posteriors at words for which the predicate is nonzero.

[9]We used the Wall-Street Journal (WSJ) portion of the Penn Treebank— sections 0–21 for training, sections 22–24 for testing.

[10]The entropy was normalized by the number of occurrences of the word.




(a) Labeling examples (b) Labeling word types

*Figure 5.* On part-of-speech tagging, choosing measurements based on entropy outperforms choosing randomly and choosing based on frequency.

run. Figure 5 shows the results: On both types of measurements, entropy outperforms choosing words at random. A simple baseline which chooses the most frequent words underperforms even random, presumably due to lack of diversity.

## 6. Conclusion

Our ultimate goal is "efficient learning"—narrowing in on the desired model with as little human effort as possible, whether it be by labeling examples or specifying constraints. Measurement-based learning allows us to integrate all of these in a coherent way. Furthermore, it is the first framework to directly target our ultimate goal by quantifying what it means to learn efficiently.

**Acknowledgments** We thank Zoubin Ghahramani for helpful comments. We wish to acknowledge support from MURI Grant N00014-06-1-0734.

## References

Borwein, J. M., & Zhu, Q. J. (2005). *Techniques of variational analysis*. Springer.

Chaloner, K., & Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science*, *10*, 273–304.

Chang, M., Ratinov, L., & Roth, D. (2007). Guiding semi-supervision with constraint-driven learning. *Association for Computational Linguistics (ACL)* (pp. 280–287).

Druck, G., Mann, G., & McCallum, A. (2008). Learning from labeled features using generalized expectation criteria. *ACM Special Interest Group on Information Retreival (SIGIR)* (pp. 595–602).

Dudík, M., Phillips, S. J., & Schapire, R. E. (2007). Maximum entropy density estimation. *Journal of Machine Learning Research*, *8*, 1217–1260.

Graça, J., Ganchev, K., & Taskar, B. (2008). Expectation maximization and posterior constraints. *Advances in Neural Information Processing Systems (NIPS)* (pp. 569–576).

Haghighi, A., & Klein, D. (2006). Prototype-driven learning for sequence models. *North American Association for Computational Linguistics (NAACL)* (pp. 320–327).

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling data. *International Conference on Machine Learning (ICML)* (pp. 282–289).

Mann, G., & McCallum, A. (2007). Simple, robust, scalable semi-supervised learning via expectation regularization. *International Conference on Machine Learning (ICML)* (pp. 593–600).

Mann, G., & McCallum, A. (2008). Generalized expectation criteria for semi-supervised learning of conditional random fields. *Human Language Technology and Association for Computational Linguistics (HLT/ACL)* (pp. 870–878).

Quadrianto, N., Smola, A. J., Caetano, T. S., & Le, Q. V. (2008). Estimating labels from label proportions. *International Conference on Machine Learning (ICML)* (pp. 776–783).

Roy, N., & McCallum, A. (2001). Toward optimal active learning through sampling estimation of error reduction. *International Conference on Machine Learning (ICML)* (pp. 441–448).

Seeger, M., & Nickisch, H. (2008). Compressed sensing and Bayesian experimental design. *International Conference on Machine Learning (ICML)* (pp. 912–919).

## A. Derivation of (12)

Let $f(q) = -H(q(Y)) - \langle E_{q(Y)}[\sum_{i=1}^n \phi(X_i, Y_i)], \theta \rangle$, $g(u) = h_\sigma(u - \tau)$, and $A(q) = E_{q(Y)}[\sigma^X(Y)]$.[11] Minimization of (9) with respect to $q$ is equivalent to minimization of $f(q) + g(Aq) + \text{constant}$. By strong duality (Theorem 4.4.3 of Borwein and Zhu (2005)), $\inf_q \{f(q) + g(Aq)\} = \sup_\beta \{-f^*(A^*\beta) - g^*(-\beta)\}$. The conjugate functions[12] are as follows: $f^*(A^*\beta) = \log \int e^{\langle \sigma^X(y), \beta \rangle + \langle \sum_{i=1}^n \phi(X_i, y_i), \theta \rangle} dy = \sum_{i=1}^n B(\beta, \theta; X_i)$ and $-g^*(-\beta) = \langle \tau, \beta \rangle - h^*_\sigma(\beta)$. Perform algebra to obtain (12).

[11] $\mathbf{W}[a] = \begin{cases} 0 & \text{if } a = \text{true} \\ \infty & \text{otherwise.} \end{cases}$

[12] The conjugate of $g(u)$ is $g^*(\beta) = \sup_u \{\langle u, \beta \rangle - g(u)\}$.