Optimal Information Gathering

--Bayesian Approaches

Tingfan Wu
Does Tingfan know SVM?

Q: What's SVM?

A: SVM is ....

It seems he knows ...

Tingfan

Committees

Acquiring information

an active information gathering
Quantification of Uncertainty

- Shannon Information Entropy
  \[ H[X] = E[-\log p(X)] = -\int p(x) \log p(x) dx \]
- Variance
  \[ \Sigma_X = Var[X] = E[(X - E[X])(X - E[X])^T] \]
- Bridge the two - Gaussian RV \( X \)
  \[ H[X] = \frac{1}{2} \log(2\pi e) + \frac{1}{2} \log \det \Sigma_X \]
More on Entropy

- conditional entropy of $X$ given $y$
  \[ H[X|y] = E[- \log p(X|y)] = - \int_x p(x|y) \log p(x|y) \, dx \]

- conditional entropy of $X$ given RV $Y$
  \[ H[X|Y] = - \int p(y) H[X|y] \, dy \]

- example
  \[ H[\Theta|Y, x] = \int_y p(y|x) H[\theta|y, x] \, dy \]
Information Gain as Uncertainty Reduction

Definition

\[ \Theta \overset{def}{=} \text{Tingfan knows SVM} \]
\[ X \overset{def}{=} \text{committee query} \]
\[ Y \overset{def}{=} \text{tingfan’s answer} \]

Information Gain (got answer \( y \))

\[ I(\theta, y|x) = H[\theta] - H[\theta|y, x] \]

initial uncertainty \quad uncertainty after query averaged over answers

Expected Information Gain (expecting answer \( Y \))

\[ I(\theta, Y|x) = H[\Theta] - H[\Theta|Y, x] \]
Ask Informative Question!

Q: What's your name?

A: My name is Ting.

still no clue

Tingfan

Committees
Information Maximization

key definition

$$\max_x I[\theta, Y|x] = \max_x H[\theta] - \int p(y|x) H[\Theta|y, x] dy$$

Bayesian Posterior Inference

$$p(\theta|y, x) \propto p(\theta)p(y|\theta, x)$$
Equivalent Formulations

- maximize information gain $\max_x I(\Theta; Y|x)$
- minimize posterior entropy $\min_x H[\Theta|x, Y]$
- maximize expected KL-divergence $\max_x \mathbb{E}[D_{KL}(p(\theta)||p(\theta|x, Y))]$
Active Learning

In a real-world setting, these instances may be available but their labels usually are not. For a given sequence, we consider different image reconstructions, and to draw items from the unlabeled pool, we correct the raw data. We simulate all further measurements under different sequences using a Cartesian grid.

Sensor Placement

Active Sensing

Visual Search

Social Interaction

Experiment Design

Posture-Specific Performance

Most misclassified labels were the ones which were most frequently confused, which is also intuitive. As the reader might have noticed, the cross-validation accuracies reported in the previous section were obtained with 31 sensors, while we deemed that the 4-shot results, given that 31 sensors were used, were obtained with 31 sensors.

Active Learning

Social Interaction

Experiment Design

OIG: An Ubiquitous Problem
OIG Property 1: **minimizing objectives**

- **model uncertainty**
  \[
  \min_x H[\Theta|x, Y]
  \]

- **predictive uncertainty**
  given \( q(x') \)
  \[
  H[Y'|x, Y, x'] = \min_x \int q(x') H[Y'|x, Y, x'] dx'
  \]
  average over \( q(x') \)
OIG Property 2: **Query Batch Size**

**Batch Size = #actions before receiving answers**

**Batch Size = 1**

1 query, 1 answer

\[
\max_x I[\theta, Y|x] = \max_x H[\theta] - \int p(y|x) H[\Theta|y, x] dy
\]

**Batch Size = 2**

2 queries, 2 answers
OIG Property 3: **Horizon**

How many questions can I ask? Plan accordingly

**BatchSize=1**

**Horizon=1**

**Greedy**

\[
X_{t+1} = f(Y_t; \eta)
\]

\[
\begin{align*}
\max_{x} I[\theta, Y|x] & = \max_{x} H[\theta] - \int p(y|x) H[\Theta|y, x] dy
\end{align*}
\]
Belief Update may not have analytic update

Integration over possible $Y$, high-dimension, long-horizon intractable

Optimize $x$: discrete pool, continuous space, or policy?
Linear Model

POMDP

Non-Linear Model

GLM (ex. Logistic Regression)

Linear Model

ex. Linear Gaussian Model
Linear-Gaussian Model

- Linear model, Gaussian prior
  
  \[ Y = \theta^T X + z, \quad \begin{cases} 
  \theta \sim \mathcal{N}(\theta_0, \Sigma_0), \\
  z \sim \mathcal{N}(0, \sigma_z) 
  \end{cases} \]

- Likelihood (also a Gaussian)
  
  \[ p(y|\theta, x) = \mathcal{N}(y|\theta^T x, \sigma_z) \]

- Posterior distribution (still Gaussian!)

\[
\begin{align*}
  p(\theta|y, x) &\propto p(\theta)p(y|\theta, x) \\
  p(\theta|y, x) &\propto \mathcal{N} \left( \theta \left| \frac{\Sigma_0^{-1} \theta_0 + \frac{x y}{\sigma_z^2}}{\Sigma_0^{-1} + \frac{x x^T}{\sigma_z^2}}, \left( \Sigma_0^{-1} + \frac{x x^T}{\sigma_z^2} \right)^{-1} \right. \right) \\
  \text{least square indep. from } y
\end{align*}
\]
Two Objectives

• min model uncertainty
• min predictive variance
Min Model Entropy

Integration
\[
\int_y p(y|x) H[\theta|y, x] dy = \int_y \log \det \left( \Sigma_0^{-1} + \frac{xx^T}{\sigma_z^2} \right)^{-1} dy + C
\]
\[
= - \log \det(\Sigma_0^{-1}) - \log \left( 1 + \frac{x^T\Sigma_0 x}{\sigma_z^2} \right) + C
\]

Optimization
\[
x^* = \arg \max_x x^T \Sigma_0 x, \text{ s.t. } ||x|| \leq M, \text{ power constraint}
\]

\[x^*\text{ parallel to eigenvector with } \lambda_{max}(\Sigma_0)\]
Visualization: Posterior Belief

\[ \theta \sim \mathcal{N}(\theta_0, \Sigma_0) \]

\[ x_1 = \arg \max_x x^T \Sigma_0 x \]

\[ p(y|\theta, x) = \mathcal{N}(y, \theta^T x_1, \sigma_z^2) \]

\[ p(\theta|x_1, y_1) \sim \mathcal{N}(\mu_1, \Sigma_1) \]

\[ x_2 = \arg \max_x x^T \Sigma_1 x \]

\[ p(\theta|x_1:2, x_1:2) \sim \mathcal{N}(\mu_2, \Sigma_2) \]
Sequential Infomax Queries

**Infomax**

\[ \theta = (0.8, 0.4) \]

**Infomax**

\[ \theta = (-0.5, -1.5) \]

**random queries**

\[ \theta = (-0.5, -1.5) \]

Variance indep. obs

Infomax better than random
Min Predictive Variance

(cf. min model entropy)

- belief update remains the same
  - predictive variance

\[ \text{Var}[y' | x', x, y] = \text{Var}[\theta^T x' | x, y] = x'^T \Sigma_{\theta|x,y} x' \]

- distribution \( q(x') \) of future question

\[
\min_x \int_{x'} q(x') \int_y p(y|x) x'^T \Sigma_{\theta|x,y} x' \, dy \, dx' = \min_x \text{tr} [\Sigma_{\theta|x,y} \Sigma_{x'}] \\
\text{indep. from } y
\]

\[ \text{reduce to } \min_x \text{tr} [\Sigma_{\theta|x,y}], \text{ if } \Sigma_{x'} = I \]

gradient descent (continuous) or \( O(n) \) iteration (discrete)
Alphabetical Optimality

Batch Size: $1 \rightarrow n$

- **D-opt** - minimize entropy of model
  \[
  \min_{x_{1:n}} H[\Theta|Y_{1:n}, x_{1:n}] = \min_{x_{1:n}} \log \det \Sigma_n
  \]

- **A-opt** - minimize variance of prediction
  \[
  \min_{x_{1:n}} \text{Var}[Y'|X', Y_{1:n}, x_{1:n}] = \min_{x_{1:n}} \text{tr}(\Sigma_n \Sigma_{x'})
  \]

A connection to classical experiment design

Integer programming relaxed to convex opt.
samples more closely around the origin. In fact, the sampling density as a function of spatial frequency has high demands on image resolution and processing time, which come with this application. We met these demands in our method.

Discussion

We believe that the main improvements are not specific to the object the sequence was optimized for. The results in Table 6 mean that in practice, randomized sequence scans are much like a gamble. The large variance across trials in Table 6 and some reconstructions vslice 8 are shown in Figure 7.

We report the performance of random designs, and to our surprise, the large variance across trials in Table 6 especially with small numbers of shots. We believe this is due to the fact that random sequences are not designed for real images, which may have artifacts or be undersampled.

We simulate all further measurements under different sequences using grid filtering. From the corrected raw data, we simulate all further measurements under different sequences. This seems small, given that the large variance across trials in Table 6 especially with small numbers of shots.

We use the sequential method, dropping this restriction disfavors equispaced designs. The spacing is more regular than is typically the case. On the other hand, the spacing is more regular than is typically the case. Short of modeling a sequence, we optimally for the Nyquist spiral by FFT, whose phase probing echoes per excitation, train of refocusing pulses, each phase encoded differently. Slices are two of them as in [45].

Interleaved outgoing corrections are judged below, is not needed. From the corrected raw data, we simulate all further measurements under different sequences. This seems small, given that the large variance across trials in Table 6 especially with small numbers of shots.

We confine ourselves to offset angles in disregard of the approximate real-valuedness of field of view FOV and to drawing by FFT, whose phase probing echoes per excitation, train of refocusing pulses, each phase encoded differently. Slices are two of them as in [45].

We optimize the sequence during the measurement, and our method are the same. We implemented requires about 50 hours on a single standard computer.

We report the performance of random designs, and to our surprise, the large variance across trials in Table 6 especially with small numbers of shots.

We use the sequential method, dropping this restriction disfavors equispaced designs. We use the sequential method, dropping this restriction disfavors equispaced designs.
Nonlinear Model

POMDP

Non-Linear Model
  ex GLM (Logistic Regression)

  Linear Model

  Gaussian
Generalized Linear Models

- Nonlinear inverse link function $g(.)$
  \[ Y \sim \text{exp. family distribution}, \ E[Y] = g(\Theta^T X), \]

- Logistic Regression (LR) \quad \quad g(\rho) = \sigma(\rho) = 1/(1 + e^{-\rho})
  \[ p(Y = 1|\theta, x) = \sigma(\theta^T x), \ Y \in \{0, 1\} \]

- LR $\sim$ many important ML algorithms
  ★ SVM - hinge vs sigmoid loss
  ★ Adaboost - sequential logistic regression

[Lewi et. al. 08]
Belief Update

A 2D example

Gaussian prior | logistic log-likelihood | true posterior | approximate posterior

\[ \log p(\theta) \quad \log p(y_t|\theta, x_t) \quad \log p(\theta|x_{1:t}, y_{1:t}) \]

\[ \mathcal{N}(\theta|\mu_{t-1}, \Sigma_{t-1}) \quad \mathcal{N}(\theta|\mu_t, \Sigma_t) \]

Gaussian approximation is asymptotically accurate

[Paninski 2005]
Belief Approximation Methods

- Discretization - flexible, doesn’t scale
- Laplace - point estimation the posterior mode
- Moment Matching -- Expectation Propagation

[Minka 2000]
Efficient Laplace Belief Update
in d-dimensional place

• Finding the mode - line search

\[ \mu_t(\Delta_t) = \mu_{t-1} + \Delta_t \Sigma_{t-1} x_t, \Delta_t \in \mathbb{R} \]

\[ \mu_t = \arg \max_{\mu_t(\Delta_t)} p(\mu_t | x_t, Y_t) \]

• Curvature at mode

\[ \Sigma_t(y_t) = - \left( \frac{\partial^2 \log p(\theta | \mu_t, C_t)}{\partial \theta \partial \theta^T} \right)^{-1} \bigg|_{\theta=\mu_t y_t} \] (original)

\[ \Sigma_t^{-1}(y_t) = \Sigma_{t-1}^{-1} - \sigma(\mu_t(y_t)^T x_t)\sigma(-\mu_t(y_t)^T x_t)x_t x_t^T \]

\[ \mathcal{O}(d^2) \text{ rank-1 update in (variance inverse)} \]
Integration and Optimization

\[ H[\theta | \mu_t(Y_t), \Sigma_t(Y_t)] = \sum_{y_t \in \{0,1\}} p(y_t | x_t, \Sigma_{t-1}, \mu_{t-1}) H[\theta | \mu_t(y_t), \Sigma_t(y_t)] \]

\[ \max_{x_t} \sum_{y \in \{0,1\}} \underbrace{\sigma(\mu_t(y_t)^T x_t) \sigma(-\mu_t(y_t)^T x_t)}_{\text{predictive uncertainty}} \underbrace{x_t^T \Sigma_t(y_t) x_t}_{\text{model uncertainty}} \]

The integrals depend on obs. \( y \).

Predictive uncertainty is represented by the \( \Sigma_t \) ellipse, and model uncertainty is represented by the \( \lambda_{\max}(\Sigma_{t-1}) \) vector.

Logistic distribution
Active Learning: pick data to label

- select from a pool, $O(n)$

Strategy: Query labels from informative examples

![Diagram showing the active learning process](image)

- learn a model
- update model
- labeled training set
- unlabeled pool $U$
- oracle (e.g., human annotator)
- select queries

Figure 1: The pool-based active learning cycle.

There are several scenarios in which active learners may pose queries, and there are also several different query strategies that have been used to decide which instances are most informative. In this section, I present two illustrative examples in the pool-based active learning setting: in which queries are selected from a large pool of unlabeled instances $U$, using an uncertainty sampling query strategy, which selects the instance in the pool about which the model is least certain how to label.

Sections 2 and 3 describe all the active learning scenarios and query strategy frameworks in more detail.

40% saving

Performance on Smile Detection Task (logistic model)
Partially Observable Markov Decision Process

POMDP

Non-Linear Model

Linear Model

flexible $\rightarrow$ hard $\rightarrow$ approximations
Infomax POMDP

Partially Observable Markov Decision Process

\[ \theta \text{ can change, spontaneously or/and caused by } X \]
Policy Representation

Re-cap

\[ x^* = \arg \max_x x^T \Sigma_0 x \]  
linear models (bs=1)

\[ x^* = \arg \max_x \text{tr}(\Sigma_\theta|x,y \Sigma'_x) \]  
bs=n

\[ x^*_{1:n} = \arg \max_{x_{1:n}} \det(\Sigma_\theta|x,y) \]  
non-linear

\[ x^* = \arg \max_{x_t} \sum_{y \in \{0,1\}} \sigma(\mu_t(y_t)^T x_t)\sigma(-\mu_t(y_t)^T x_t)x_t^T \Sigma_t(y_t)x_t \]

next question(s) = \( \pi(\theta|x_{1:t}, y_{1:t}) \)

suff. stat. of belief

\( \mu_t, \Sigma_t \)  
Gaussian

\( b^i_t = p(\theta_t = i) \)  
belief states

flexible discretization
Belief Update

\[ b_t^i \propto p(Y|\theta_t = i, x_{t-1}) \sum_{j=1}^{N} p(\theta_t = i|x_{t-1}, \theta_{t-1} = j)b_{t-1}^j \]

(posterior belief) = (likelihood) (transition) (prior belief)

* Special Case: static parameter

\[ p(\theta_t = i|x_{t-1}, \theta_{t-1} = j) = 1 \text{ if } i = j \]
Average Reward by Integration

- **Infomax POMDP reward:** negative belief entropy

  \[
  R_t(b_t, x_t) = \sum_i b_t^i \log b_t^i
  \]

  encourage low-uncertainty

  \[
  -H[\theta|x_{1:t}, y_{1:t}]
  \]

- **Parameterized policy, a popular approximation**

  \[
  x_t = \pi(b_t; \eta)
  \]

- **Often Horizon=\(T>1\), evaluate Policy by Monte-Carlo**

  \[
  E \left[ \sum_{t=1}^{T} R(B_t, X_t) \right] \approx \frac{1}{N} \sum_{x_{1:T} \sim \pi(b_{1:T})} \sum_{t=1}^{T} \log \pi(x_i|b_i; \eta) \sum_{t=1}^{T} R(b_t, x_t)
  \]
Gradient Ascent Optimization

Update rule
\[ \eta^{t+1} = \eta^t + \alpha \nabla_{\eta} E \left[ \sum_{t=1}^{T} R(B_t, X_t) \mid \eta \right] \]
\[ \approx \sum_{j=1}^{N} \left( \sum_{t=1}^{T} \nabla \log \pi(x_t \mid b_t; \eta) \sum_{i=t}^{T} r(b_i) \right) \]
reward weighted gradient

Find a policy maximizing infomax reward
\[ \arg \max_{\eta} E \left[ \sum_{t=1}^{T} R(B_t, X_t) \mid \eta \right] \]

Avoid optimizing over \( x_{1:n} \) by finding good \( \eta \) in policy space
Social Robot

Track peoples’ (location, expression) that change with time

Multimodal (face-smile detector, audio localizer), various-FoV

Learned Policy emerging human-like behavior

[Fasel et. al.]
Infomax for Object Detector

Typical Scanning Detections

Infomax Digital Retina

update belief from windows across scales and locations
Digital Eye in Action

[Butko 2009]
Submodularity
Optimal Sensor Deployment

The temperatures \((\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) \sim \mathcal{GP}(\mu, \Sigma), \Sigma \text{ known}\)

\[
V = \{1, 2, 3, 4, 5\} \\
A = \{2, 3\}
\]

\(X_i: \text{pick the point to measure}\)  \\
\(Y_i: \text{noiseless obs. of } \theta\)

Infomax: \[
\max_{A \subset V, |A| = k} I(\theta, Y_A | x_A)
\]
Solving the Problem

• **Posterior Belief Distribution**
  \[ \Sigma_{V|A} = \Sigma_{VV} - \Sigma_{VA} \Sigma_{AA}^{-1} \Sigma_{AV} \]

• **Integration**
  \[
  \int p(y_A|x_A) H[\theta|y_A, x_A] dy_A = \log \det \Sigma_{V|A}
  \]

• **Find Optimal A**
  \[
  \max_{A \subset \mathcal{V}, |A| = k} I(\theta, Y_A|x_A)
  \]

  Let \( F(A) = I(V, A) = I(V \setminus A, A) \)

  \( F \) is submodular and increasing in \(|A|\),
greedy is approximate optimal

  Assume \( 2|A| < |\mathcal{V}| \)
Submodular Optimization and Optimality

\[
\text{Function } \hat{A} = \text{Greedy}(F, k) \\
\hat{A}_0 = \emptyset \\
\text{for } j = 1 \ldots k \text{ do} \\
\quad \hat{A}_j \leftarrow \hat{A}_{j-1} \cup \arg \max_{X \in \mathcal{V} \setminus A_{j-1}} F(A_{j-1} \cup X) \\
\text{end} \\
\text{return } \hat{A}_j
\]

Theorem: If \( F \) is submodular and increasing

\[
F(\hat{A}) \geq 1 - (1 - \frac{1}{k})^k \geq (1 - \frac{1}{e}) \max_{|A|=k} F(A).
\]
Pressure Sensors on a Chair

Save 98% cost for similar 10-posture classification accuracy
Conclusion

- Exploiting known structures in unobserved RVs
- Active > Passive information gathering
- Efficient approximations may be necessary

Have you gotten the “information” you need?