Knowledge Rerrangement

Empirical Bayes

Method

"Learning the experience from others"

Examples & Approaches[1]

- Robbins' Formula, The Missing Species Problem, James-Stein Estimator, False Discovery Rate......
- f-modeling & g-modeling

Recent research on EB

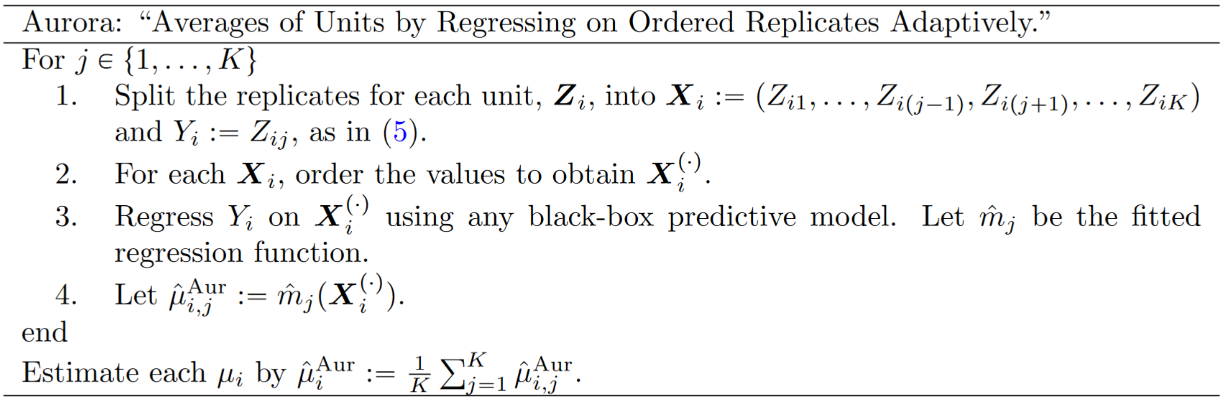
EB method when known F and unknown G

- General maximum likelihood empirical Bayes estimation of normal means[2]
- Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means[3]
- High dimensional exponential family estimation via empirical Bayes[4]
- On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising[5]

EB method with unknown F and G

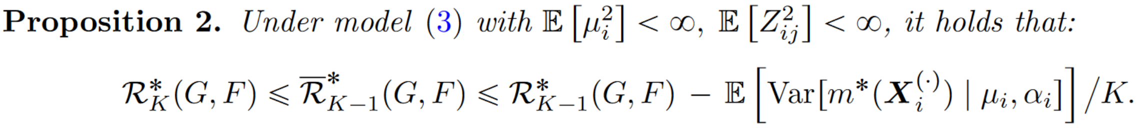
Order statistic regression on replicated data[6]

- Driven Force: Replication makes it possible to estimate μ_i with no assumptions on F and G with the aim of matching the risk of Bayes rule
- Target: Point estimation of the posterior mean $\mathbb{E}_F[Z_{ij}|\mu_i, \alpha_i]$
- Key Method: The key insight is that the conditional mean $\mathbb{E}_{F,G}[Z_{ij}|X_i]$ is (almost surely) identical to the posterior mean $\mathbb{E}_{F,G}[\mu_i|X_i]$, which represents that Bayes rule could be estimated via $\mathbb{E}_{F,G}[\mu_i|X_i]$ and simply regress Y_i on X_i under any black-box predictive model.
- **Self-interpretation:** Learning the experience from replicates so that we don't need stronger assumption of prior and marginal likelihood when our target is to estimate the posterior mean, which regression (no matter black-box or not) method with order statistics under exchangeable sampling, Aurora is an effective approach that could draw nearly Bayes risk.
- Expansion to high-dimensional situations is natural. Two approaches: mean or the order statistics(proposed cuz the mean may not be the sufficient statistics for (μ_i, α_i)); Average all to reduce the variance.
- Algorithms:



```
Aurora: "Averages of Units by Regressing on Ordered Replicates Adaptively."
For $j \in \{1, \ldots, K\}$
    1. Split the replicates for each unit, $\boldsymbol{Z}_i$, into $\boldsymbol{X}_i := (Z_{i1}, \ldots, Z_{i(j-1)}, Z_{i(j+1)}, \ldots, Z_{iK})$
       and $Y_i := Z_{ij}$, as in (5).
    2. For each $\boldsymbol{X}_i$, order the values to obtain $\boldsymbol{X}_i^{(\cdot)}$.
    3. Regress $Y_i$ on $\boldsymbol{X}_i^{(\cdot)}$ using any black-box predictive model. Let $\hat{m}_j$ be the fitted
       regression function.
    4. Let $\hat{\mu}_{i,j}^{\text{Aur}} := \hat{m}_j(\boldsymbol{X}_i^{(\cdot)})$.
end
Estimate each $\mu_i$ by $\hat{\mu}_i^{\text{Aur}} := \frac{1}{K}\sum_{j=1}^{K} \hat{\mu}_{i,j}^{\text{Aur}}$.
```

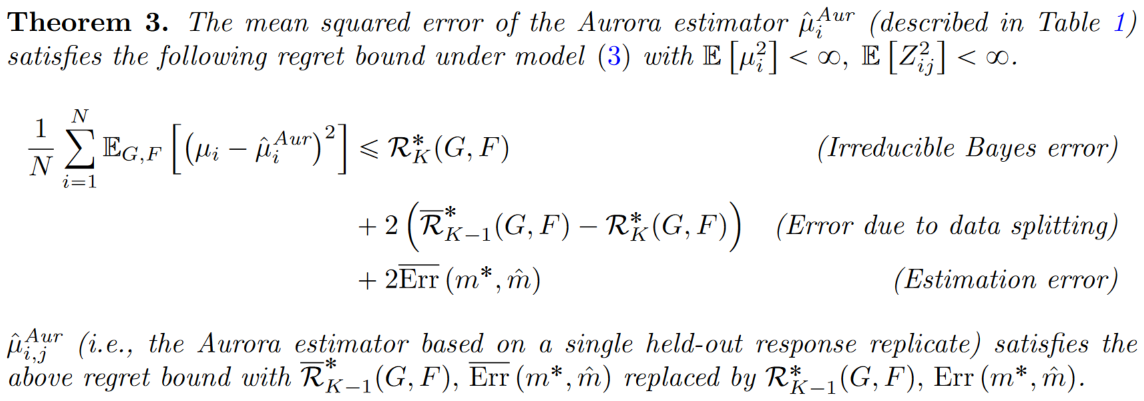
- Properties:
    1. Bayes Risk of Aurora



Proposition 2. *Under model* (3) *with* $\mathbb{E}\left[\mu_i^2\right] < \infty$, $\mathbb{E}\left[Z_{ij}^2\right] < \infty$, *it holds that:*

$$\mathcal{R}_K^*(G,F) \leqslant \overline{\mathcal{R}}_{K-1}^*(G,F) \leqslant \mathcal{R}_{K-1}^*(G,F) \; - \; \mathbb{E}\left[\mathrm{Var}[m^*(\boldsymbol{X}_i^{(\cdot)}) \mid \mu_i, \alpha_i]\right]/K.$$

2. Regret bound and decomposition of Aurora

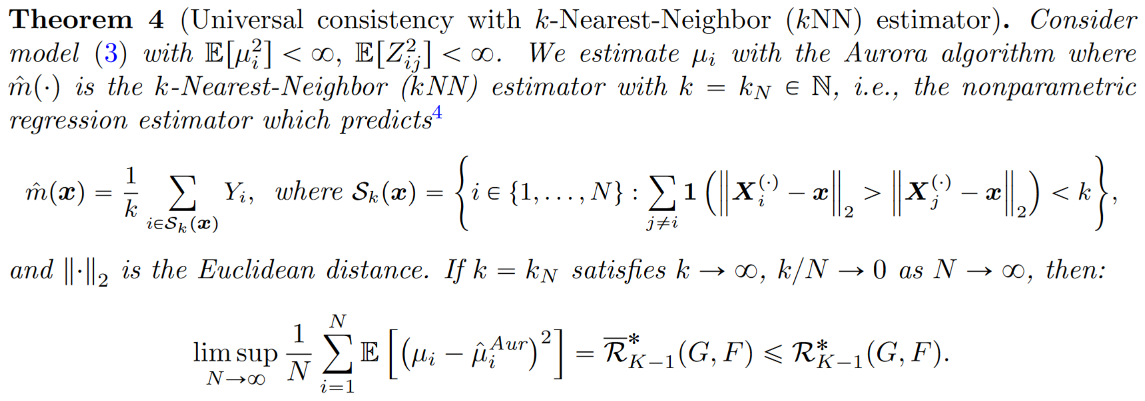


Theorem 3. *The mean squared error of the Aurora estimator $\hat{\mu}_i^{Aur}$ (described in Table 1) satisfies the following regret bound under model (3) with $\mathbb{E}\left[\mu_i^2\right] < \infty$, $\mathbb{E}\left[Z_{ij}^2\right] < \infty$.*

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{G,F}\left[\left(\mu_i - \hat{\mu}_i^{Aur}\right)^2\right] \leqslant\ & \mathcal{R}_K^*(G,F) && \textit{(Irreducible Bayes error)} \\ & + 2\left(\overline{\mathcal{R}}_{K-1}^*(G,F) - \mathcal{R}_K^*(G,F)\right) && \textit{(Error due to data splitting)} \\ & + 2\overline{\mathrm{Err}}\left(m^*, \hat{m}\right) && \textit{(Estimation error)} \end{aligned}$$

$\hat{\mu}_{i,j}^{Aur}$ (i.e., the Aurora estimator based on a single held-out response replicate) satisfies the above regret bound with $\overline{\mathcal{R}}_{K-1}^*(G,F)$, $\overline{\mathrm{Err}}\left(m^*, \hat{m}\right)$ replaced by $\mathcal{R}_{K-1}^*(G,F)$, $\mathrm{Err}\left(m^*, \hat{m}\right)$.

3. Aurora with KNN



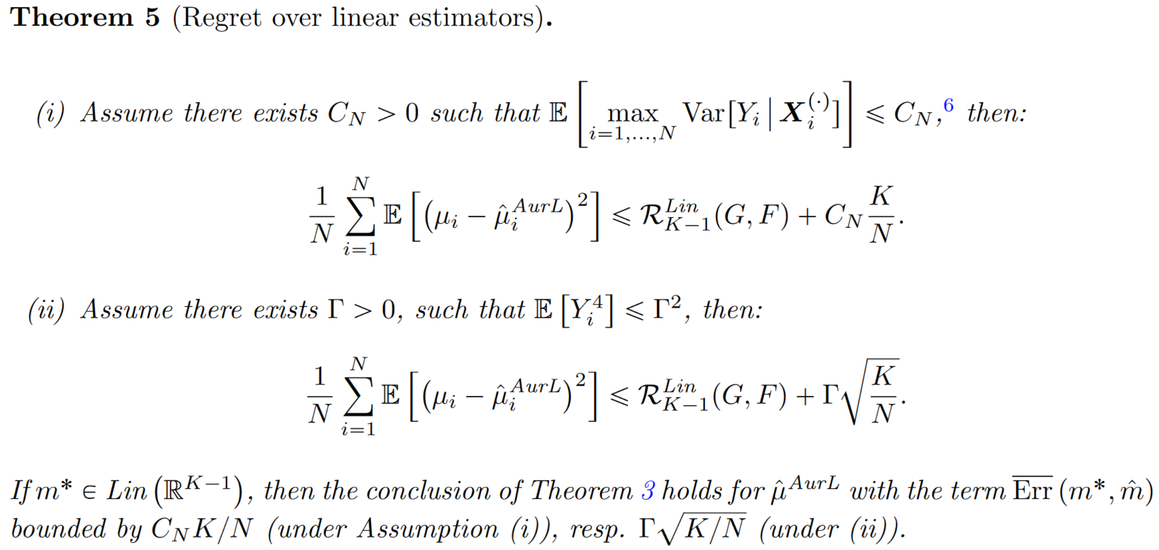
Theorem 4 (Universal consistency with k-Nearest-Neighbor (kNN) estimator)**.** *Consider model (3) with $\mathbb{E}[\mu_i^2] < \infty$, $\mathbb{E}[Z_{ij}^2] < \infty$. We estimate μ_i with the Aurora algorithm where $\hat{m}(\cdot)$ is the k-Nearest-Neighbor (kNN) estimator with $k = k_N \in \mathbb{N}$, i.e., the nonparametric regression estimator which predicts*[4]

$$\hat{m}(\boldsymbol{x}) = \frac{1}{k}\sum_{i \in \mathcal{S}_k(\boldsymbol{x})} Y_i, \quad \textit{where } \mathcal{S}_k(\boldsymbol{x}) = \left\{ i \in \{1,\ldots,N\} : \sum_{j \neq i} \mathbf{1}\left(\left\|\boldsymbol{X}_i^{(\cdot)} - \boldsymbol{x}\right\|_2 > \left\|\boldsymbol{X}_j^{(\cdot)} - \boldsymbol{x}\right\|_2\right) < k \right\},$$

and $\|\cdot\|_2$ is the Euclidean distance. If $k = k_N$ satisfies $k \to \infty$, $k/N \to 0$ as $N \to \infty$, then:

$$\limsup_{N \to \infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\left(\mu_i - \hat{\mu}_i^{Aur}\right)^2\right] = \overline{\mathcal{R}}_{K-1}^*(G,F) \leqslant \mathcal{R}_{K-1}^*(G,F).$$

4. Aurora in Linear regression



Theorem 5 (Regret over linear estimators)**.**

(i) Assume there exists $C_N > 0$ such that $\mathbb{E}\left[\max_{i=1,\ldots,N} \mathrm{Var}[Y_i \mid \boldsymbol{X}_i^{(\cdot)}]\right] \leqslant C_N$,[6] *then:*

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\left(\mu_i - \hat{\mu}_i^{AurL}\right)^2\right] \leqslant \mathcal{R}_{K-1}^{Lin}(G,F) + C_N \frac{K}{N}.$$

(ii) Assume there exists $\Gamma > 0$, such that $\mathbb{E}\left[Y_i^4\right] \leqslant \Gamma^2$, then:

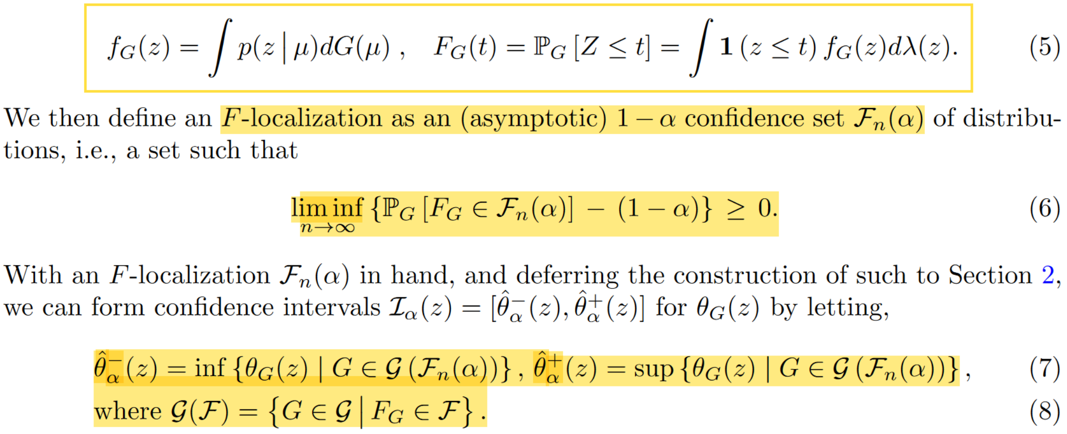
$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\left(\mu_i - \hat{\mu}_i^{AurL}\right)^2\right] \leqslant \mathcal{R}_{K-1}^{Lin}(G,F) + \Gamma\sqrt{\frac{K}{N}}.$$

If $m^* \in Lin\left(\mathbb{R}^{K-1}\right)$, then the conclusion of Theorem 3 holds for $\hat{\mu}^{AurL}$ with the term $\overline{\mathrm{Err}}\left(m^*, \hat{m}\right)$ bounded by $C_N K/N$ (under Assumption (i)), resp. $\Gamma\sqrt{K/N}$ (under (ii)).

Confidence Intervals for Nonparametric Empirical Bayes Analysis[7]

- Starting Point: The importance of measuring the uncertainty of the estimators using empirical bayes methods
- Two approaches:

1. F-localization;
   - Key idea: Construct a confidence set for the marginal distribution of Z and then determine all G ∈ G consistent with this confidence set
   - Procedures:



$$\boxed{f_G(z) = \int p(z \mid \mu) dG(\mu)\,, \quad F_G(t) = \mathbb{P}_G\left[Z \le t\right] = \int \mathbf{1}\left(z \le t\right) f_G(z) d\lambda(z).} \tag{5}$$

We then define an F-localization as an (asymptotic) $1-\alpha$ confidence set $\mathcal{F}_n(\alpha)$ of distributions, i.e., a set such that

$$\liminf_{n\to\infty} \left\{\mathbb{P}_G\left[F_G \in \mathcal{F}_n(\alpha)\right] \,-\, (1-\alpha)\right\} \,\ge\, 0. \tag{6}$$

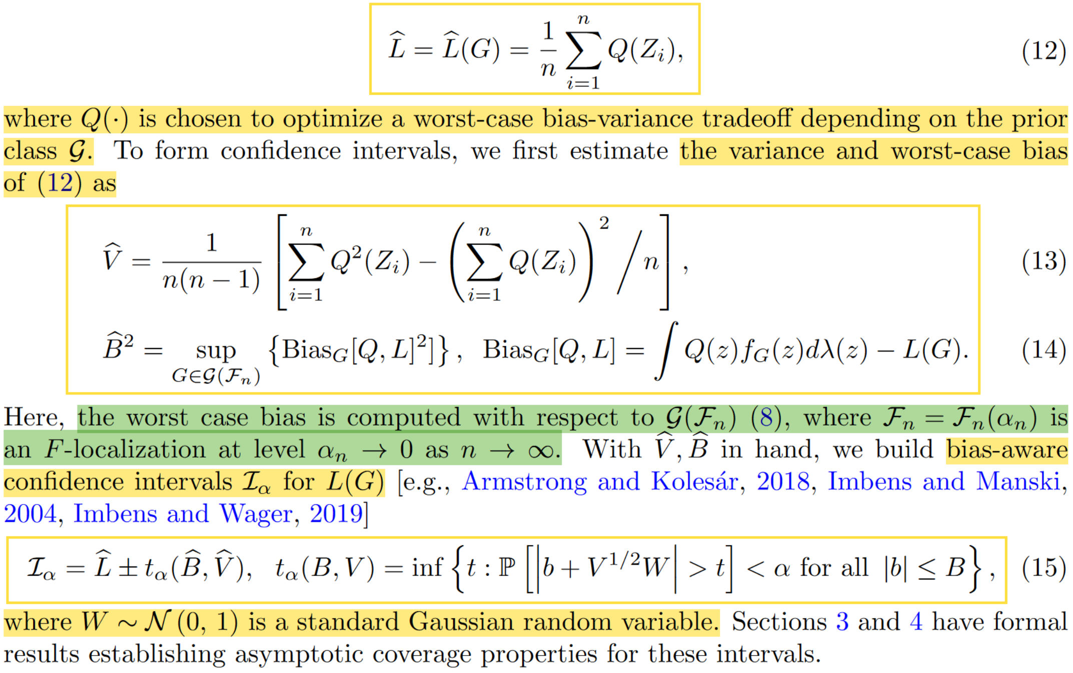
With an F-localization $\mathcal{F}_n(\alpha)$ in hand, and deferring the construction of such to Section 2, we can form confidence intervals $\mathcal{I}_\alpha(z) = [\hat{\theta}^-_\alpha(z), \hat{\theta}^+_\alpha(z)]$ for $\theta_G(z)$ by letting,

$$\hat{\theta}^-_\alpha(z) = \inf\left\{\theta_G(z) \mid G \in \mathcal{G}\left(\mathcal{F}_n(\alpha)\right)\right\}, \; \hat{\theta}^+_\alpha(z) = \sup\left\{\theta_G(z) \mid G \in \mathcal{G}\left(\mathcal{F}_n(\alpha)\right)\right\}, \tag{7}$$

$$\text{where } \mathcal{G}(\mathcal{F}) = \left\{G \in \mathcal{G} \,\middle|\, F_G \in \mathcal{F}\right\}. \tag{8}$$

   - Two common choices: Gauss-F-localization & χ^2-F-localization, one for continuous likelihood and another for categorical likelihood.

2. AMARI
   - Starting point: The posterior mean could be rewritten as a fraction of linear functionals of G
   - Turn the hypothesis of an estimator $\theta_G(Z)$ into the construction of confidence intervals for linear functionals $L(G)$. The core proposal is to estimate it as an affine estimator:



$$\widehat{L} = \widehat{L}(G) = \frac{1}{n}\sum_{i=1}^{n} Q(Z_i), \tag{12}$$

where $Q(\cdot)$ is chosen to optimize a worst-case bias-variance tradeoff depending on the prior class $\mathcal{G}$. To form confidence intervals, we first estimate the variance and worst-case bias of (12) as

$$\widehat{V} = \frac{1}{n(n-1)}\left[\sum_{i=1}^{n} Q^2(Z_i) - \left(\sum_{i=1}^{n} Q(Z_i)\right)^2 \Big/ n\right], \tag{13}$$

$$\widehat{B}^2 = \sup_{G\in\mathcal{G}(\mathcal{F}_n)} \left\{\text{Bias}_G[Q,L]^2\right\}, \quad \text{Bias}_G[Q,L] = \int Q(z) f_G(z) d\lambda(z) - L(G). \tag{14}$$

Here, the worst case bias is computed with respect to $\mathcal{G}(\mathcal{F}_n)$ (8), where $\mathcal{F}_n = \mathcal{F}_n(\alpha_n)$ is an F-localization at level $\alpha_n \to 0$ as $n \to \infty$. With $\widehat{V}, \widehat{B}$ in hand, we build bias-aware confidence intervals $\mathcal{I}_\alpha$ for $L(G)$ [e.g., Armstrong and Kolesár, 2018, Imbens and Manski, 2004, Imbens and Wager, 2019]

$$\mathcal{I}_\alpha = \widehat{L} \pm t_\alpha(\widehat{B}, \widehat{V}), \quad t_\alpha(B,V) = \inf\left\{t : \mathbb{P}\left[\left|b + V^{1/2}W\right| > t\right] < \alpha \text{ for all } |b| \leq B\right\}, \tag{15}$$

where $W \sim \mathcal{N}(0, 1)$ is a standard Gaussian random variable. Sections 3 and 4 have formal results establishing asymptotic coverage properties for these intervals.

Both two approches will finally be transformed into an optimization problem using the Charnes and Cooper transformation for linear-fractional programming.

- Difference between two approaches is that F-localization constructs a simultaneous interval while AMARI constructs pointwise interval.

Meta-Learning

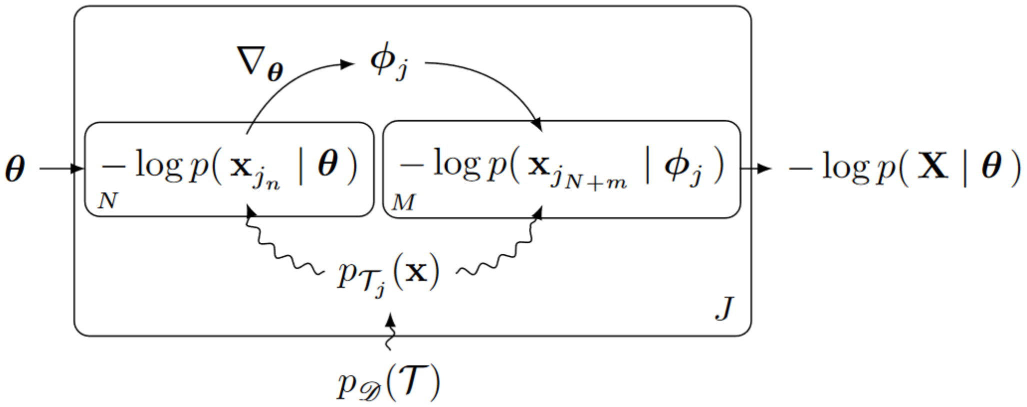
Relationship between MAML and Empirical Bayes[8]

- High requirement of generalization ability in multiple tasks → Meta-Learning
- Two methods of the construction of meta-learning:
    1. Gradient-based hyperparameter optimization(MAML as an example)
        - Provides a gradient-based meta-learning procedure that employs a single additional parameter (the meta-learning rate) and operates on the same parameter space for both meta-learning and fast adaptation.

- The objective of MAML:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{J}\sum_j \left[\frac{1}{M}\sum_m -\log p(\mathbf{x}_{j_{N+m}} \mid \underbrace{\boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \frac{1}{N}\sum_n -\log p\left(\mathbf{x}_{j_n} \mid \boldsymbol{\theta}\right))}_{\boldsymbol{\phi}_j} \right]$$

- Procedure:

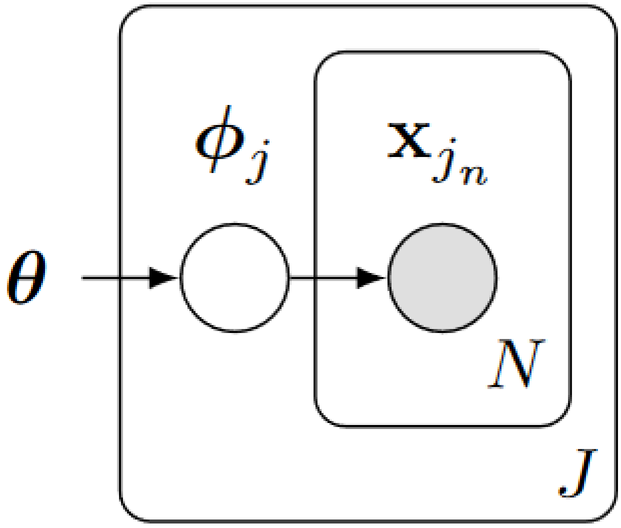

$\nabla_{\boldsymbol{\theta}}$
$\boldsymbol{\phi}_j$
$\boldsymbol{\theta}$
$-\log p(\mathbf{x}_{j_n} \mid \boldsymbol{\theta})$
N
$-\log p(\mathbf{x}_{j_{N+m}} \mid \boldsymbol{\phi}_j)$
M
$-\log p(\mathbf{X} \mid \boldsymbol{\theta})$
$p_{\mathcal{T}_j}(\mathbf{x})$
J
$p_{\mathcal{D}}(\mathcal{T})$

- In particular, in the case of meta-learning, each task-specific parameter ϕ_j is distinct from but should influence the estimation of the parameters from other tasks.

2. Probabilistic inference in a hierarchical Bayesian model
   - Objective:

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_j \left(\int p\left(\mathbf{x}_{j_1}, \ldots, \mathbf{x}_{j_N} \mid \boldsymbol{\phi}_j\right) p\left(\boldsymbol{\phi}_j \mid \boldsymbol{\theta}\right) \mathrm{d}\boldsymbol{\phi}_j \right)$$

   We maximize this objective as a function of θ and apply empirical bayes to estimate the prior parameters.

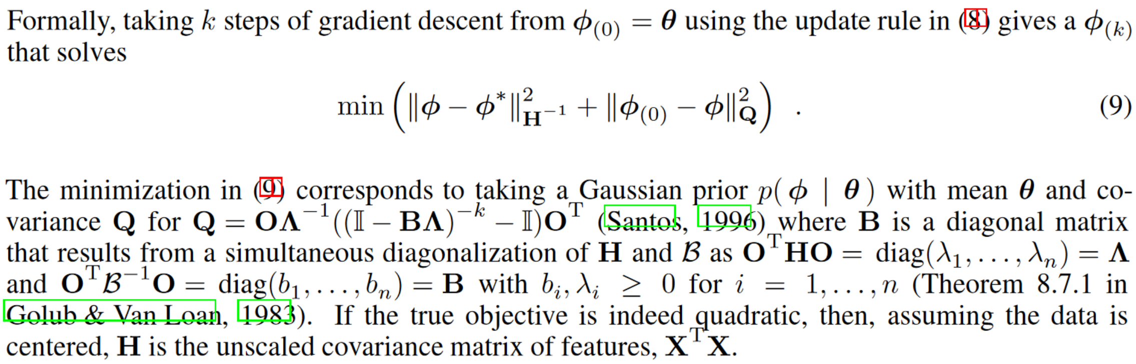


    - Procedures:

- Connection between two methods:
  1. Conclusion: MAML can be understood as empirical Bayes in a hierarchical probabilistic model!
  2. Derivation:
     - The task-specific parameter ϕ_j's exactly marginal distribution in the objective of hierarchical bayesian model is not tractable to obtain
     - Consider an approximation of that objective using $\hat{\phi}_j$:
       $$-\log p\left(\mathbf{X} \mid \boldsymbol{\theta}\right) \approx \sum_j\left[-\log p\left(\mathbf{x}_{j_{N+1}}, \ldots \mathbf{x}_{j_{N+M}} \mid \hat{\boldsymbol{\phi}}_j\right)\right]$$
  - Setting $\hat{\phi}_j=\theta+\alpha \nabla_\theta logp(x_{j_1}, \ldots, x_{j_N} \mid \theta)$ we could obtain the unscaled form of objective in gradient-based hyperparameter optimization. Trade-off between optimizing the objective and staying close to θ.
  - Consider the second-order approximation of fast adaptation: $\ell(\phi) \approx \tilde{\ell}(\phi):=\frac{1}{2}\|\phi-\phi^*\|_{\mathbf{H}^{-1}}^2+\ell\left(\phi^*\right)$
  - $\phi_{(k)}=\phi_{(k-1)}-\mathcal{B} \nabla_\phi \tilde{\ell}\left(\phi_{(k-1)}\right)$ where B denotes the curvature matrix:
    1. Reflects the Newton's method if B is diagonal

2. Incorperate the task-general information into the covariance of the fast adaptation so that reflects the interaction between task-specific parameters

- Formalization:



Formally, taking k steps of gradient descent from $\phi_{(0)} = \boldsymbol{\theta}$ using the update rule in (8) gives a $\phi_{(k)}$ that solves

$$\min\left(\|\phi - \phi^*\|^2_{\mathbf{H}^{-1}} + \|\phi_{(0)} - \phi\|^2_{\mathbf{Q}}\right) \, . \tag{9}$$

The minimization in (9) corresponds to taking a Gaussian prior $p(\phi \mid \boldsymbol{\theta})$ with mean $\boldsymbol{\theta}$ and covariance $\mathbf{Q}$ for $\mathbf{Q} = \mathbf{O}\boldsymbol{\Lambda}^{-1}((\mathbb{I} - \mathbf{B}\boldsymbol{\Lambda})^{-k} - \mathbb{I})\mathbf{O}^{\mathrm{T}}$ (Santos, 1996) where $\mathbf{B}$ is a diagonal matrix that results from a simultaneous diagonalization of $\mathbf{H}$ and $\mathcal{B}$ as $\mathbf{O}^{\mathrm{T}}\mathbf{H}\mathbf{O} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n) = \boldsymbol{\Lambda}$ and $\mathbf{O}^{\mathrm{T}}\mathcal{B}^{-1}\mathbf{O} = \mathrm{diag}(b_1, \ldots, b_n) = \mathbf{B}$ with $b_i, \lambda_i \geq 0$ for $i = 1, \ldots, n$ (Theorem 8.7.1 in Golub & Van Loan, 1983). If the true objective is indeed quadratic, then, assuming the data is centered, $\mathbf{H}$ is the unscaled covariance matrix of features, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

- Two ways to improve the perfomance of MAML:
  1. Laplace method for integration
  2. Curvature information(K-FAC)

Recommender System & MAB

Multi-armed Bandits

What do we expect?

Find strategies for Exploration-Exploitation problem.

Formulation

Notations settings

- $\mathcal{A}$: finite set of actions;
- $R_t, t \in \mathbb{Z}_{++}$: random sequence of reward vectors;
- $\mathcal{E}$: environment, a probability measure-valued random variable that takes on values in the set of all probability measures on $(\mathbb{R}^{\mathcal{A}}, \mathcal{B})$, where latter one is the Borel sigma-algebra;
- $\theta = \mathbb{E}[R_1|\mathcal{E}]$: mean rewards;
- $A_* \sim unif(argmax\theta_a)$: optimal action;

- $\mathcal{H}_t$: history at time t;
- π: agent policy;
- $H_t^\pi = (A_0^\pi, R_{1,A_0^\pi}, \ldots, A_{t-1}^\pi, R_{t,A_{t-1}^\pi})$: the history generated as an agent executes policy π by sampling each action A_t^π from $\pi(\cdot|H_t^\pi)$ and receives the resulting reward R_{t+1,A_t^π} ;
- $R_* = max_{a\in\mathcal{A}}\theta_a$: maximum reward across actions;
- $\mathcal{R}(T) = \mathbb{E}[\sum_{t=0}^{T-1}(R_* - R_{t+1,A_t})]$: cumulative Bayesian regret.

Information Ratio

- Basic version:

$$\Gamma_{\mathcal{E}} = \sup_{t\in\mathbb{Z}_+, h\in\mathcal{H}_t} \frac{\mathbb{E}[R_* - R_{t+1,A_t}|H_t = h]^2}{\mathbb{I}(\mathcal{E}; A_t, R_{t+1,A_t}|H_t = h)}.$$

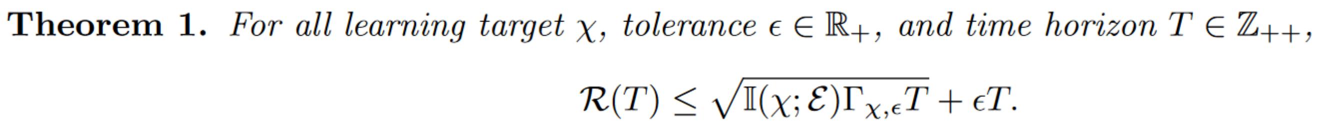
- Information ratio, defined with respect to a learning target χ, which is a random variable for which $\chi\perp H_\infty|\mathcal{E}$:

$$\Gamma_{\chi} = \sup_{t\in\mathbb{Z}_+, h\in\mathcal{H}_t} \frac{\mathbb{E}[R_* - R_{t+1,A_t}|H_t = h]^2}{\mathbb{I}(\chi; A_t, R_{t+1,A_t}|H_t = h)}.$$

- With additional tolerance ϵ:

$$\Gamma_{\chi,\epsilon} = \sup_{t\in\mathbb{Z}_+, h\in\mathcal{H}_t} \frac{\mathbb{E}[R_* - R_{t+1,A_t} - \epsilon|H_t = h]_+^2}{\mathbb{I}(\chi; A_t, R_{t+1,A_t}|H_t = h)}.$$

Regret Bound

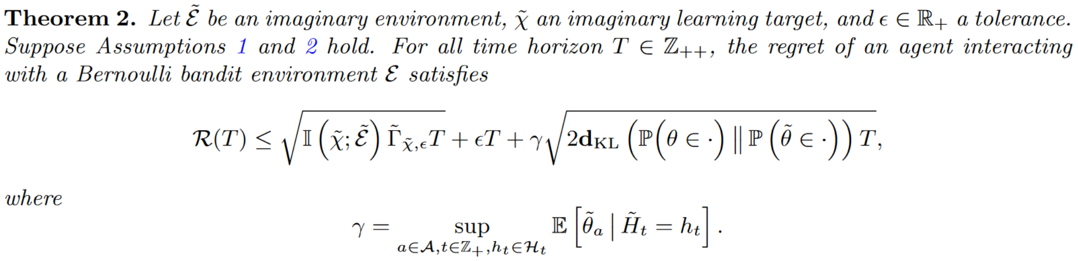


Theorem 1. *For all learning target χ, tolerance $\epsilon \in \mathbb{R}_+$, and time horizon $T \in \mathbb{Z}_{++}$,*

$$\mathcal{R}(T) \leq \sqrt{\mathbb{I}(\chi;\mathcal{E})\Gamma_{\chi,\epsilon}T} + \epsilon T.$$

Bayes in Bandits

- Original Thompson Sampling: Conjucate Prior (Beta) in Bernoulli bandit settings
- What if the distribution behind the prior and likelihood function is not conjucate(inconsistent) to the bandits settings? For example, Gaussian in Bernoulli?[9]
  - Core Method: Utilize the notion of change of measure and set assumptions about it, decompose the regret bound.
  - Procedure in Gaussian Imagination(i.e. regard the observed data from true belief as generation of Gaussian prior):
    1. Imitate the formulation above about the true environment, we formulate the notations in the imaginated environment.
    2. Setting assumptions:
       - For all $t \in \mathbb{Z}_{++}$, $\mathbb{E}\left[\mathbb{E}\left[\tilde{R}_* \mid \tilde{H}_t \leftarrow H_t\right]\right] \geq \mathbb{E}\left[R_*\right]$
       - The imaginary learning target $\tilde{\chi}$ and the imaginary mean reward $\tilde{\theta}$ are jointly Gaussian.
    3. Main theorem:



Theorem 2. *Let $\tilde{\mathcal{E}}$ be an imaginary environment, $\tilde{\chi}$ an imaginary learning target, and $\epsilon \in \mathbb{R}_+$ a tolerance. Suppose Assumptions 1 and 2 hold. For all time horizon $T \in \mathbb{Z}_{++}$, the regret of an agent interacting with a Bernoulli bandit environment $\mathcal{E}$ satisfies*

$$\mathcal{R}(T) \leq \sqrt{\mathbb{I}\left(\tilde{\chi};\tilde{\mathcal{E}}\right)\tilde{\Gamma}_{\tilde{\chi},\epsilon}T} + \epsilon T + \gamma\sqrt{2\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(\theta \in \cdot\right) \| \mathbb{P}\left(\tilde{\theta} \in \cdot\right)\right)T},$$

where

$$\gamma = \sup_{a \in \mathcal{A}, t \in \mathbb{Z}_+, h_t \in \mathcal{H}_t} \mathbb{E}\left[\tilde{\theta}_a \mid \tilde{H}_t = h_t\right].$$

    4. Brief proof of this theorem

Bayesian Linear Probit Model

Expectation Propogation[10]

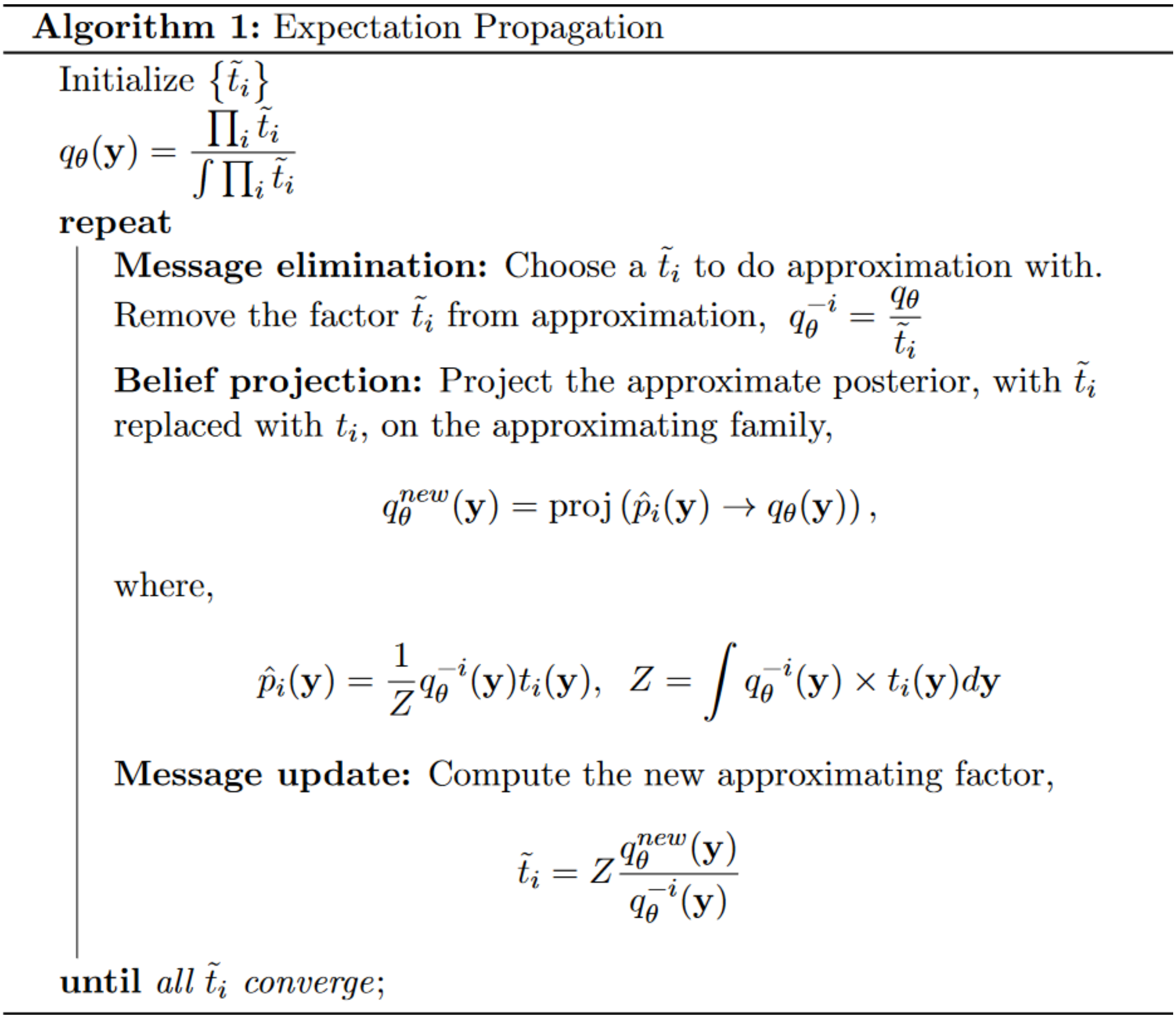
Assumed Density Filtering(ADF)

- Moment Matching, Weak Marginalization, etc
- Goal: Find an exact posterior $p(y|x)$ and its approximation $q(y)$, which could be viewed as a projection towards the another

family of distributions.
- Here we project the posterior towards the expotential family via KL-divergence
- $\nabla_\theta KL(p||q) = 0 \Rightarrow \mathbb{E}_q[\Phi(y)] = \mathbb{E}_p[\Phi(y)]$, which represent that it's enough to match their moments to minimize the KL-divergence between them.
- For a factorized distribution, the order of factors will change the final approximation of this posterior.

Expectation Propogation

- Solve the problem in ADF that the impact of the order of factors.
- Update the approximation in a recursive way.
- Algorithms:



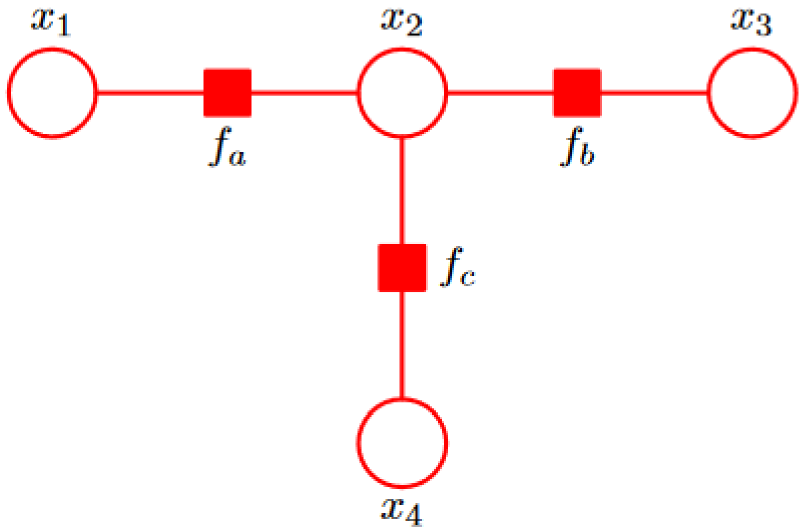
```
Algorithm 1: Expectation Propagation
  Initialize $\{\tilde{t}_i\}$
  $q_\theta(\mathbf{y}) = \frac{\prod_i \tilde{t}_i}{\int \prod_i \tilde{t}_i}$
  repeat
      Message elimination: Choose a $\tilde{t}_i$ to do approximation with.
      Remove the factor $\tilde{t}_i$ from approximation, $q_\theta^{-i} = \frac{q_\theta}{\tilde{t}_i}$
      Belief projection: Project the approximate posterior, with $\tilde{t}_i$
      replaced with $t_i$, on the approximating family,
          $q_\theta^{new}(\mathbf{y}) = \text{proj}\left(\hat{p}_i(\mathbf{y}) \rightarrow q_\theta(\mathbf{y})\right),$
      where,
          $\hat{p}_i(\mathbf{y}) = \frac{1}{Z} q_\theta^{-i}(\mathbf{y}) t_i(\mathbf{y}), \;\; Z = \int q_\theta^{-i}(\mathbf{y}) \times t_i(\mathbf{y}) d\mathbf{y}$
      Message update: Compute the new approximating factor,
          $\tilde{t}_i = Z \frac{q_\theta^{new}(\mathbf{y})}{q_\theta^{-i}(\mathbf{y})}$
  until all $\tilde{t}_i$ converge;
```

- Shortcomings: sensitive to outliers; no known convergene proof

Factor Graph & Sum Product Algorithm[11][12]

Factor Graph

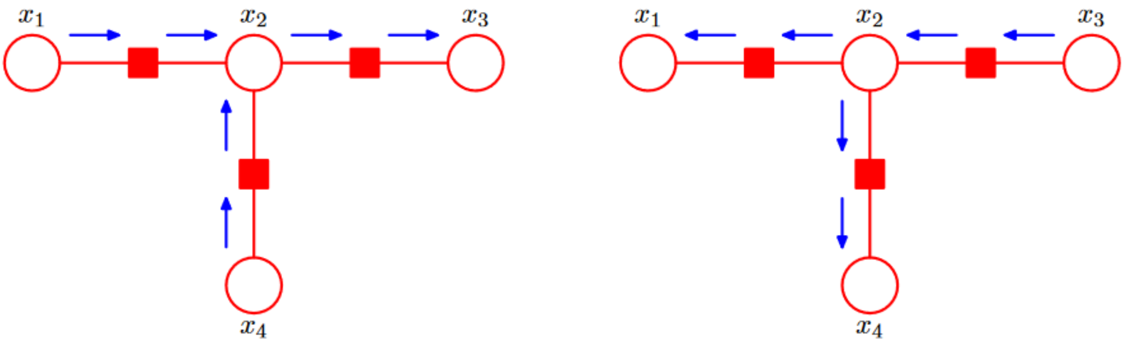
- Generization of undirected and directed graph.
- Express the dependence of variables and factors
- Factor node(rectangle point,functions)
  Variable node(circle point)
  Link(dependence between the factor and variable)



Sum-Product Algorithm

- Belief Propogation is a special case of Sum-Product Algorithm
- Goal:obtain marginal distribution of arbitrary variables;
- $p(x) = \sum_{\mathbf{X}\backslash x} p(x) = \sum_{\mathbf{X}\backslash x} [\prod_{s \in ne(x)} F_x(x, X_s)] = \prod_{s \in ne(x)} [\sum_{X_s} F_x(x, X_s)] =$
- $\mu_{f_s \to x}(x) = \sum_{x_1} \dots \sum_{x_M} f_s(x, x_1, \dots, x_M) \prod_{m \in ne(f_s)\backslash x} [\sum_{X_{xm}} G_m(x_m, X_{sm})]$
- $\mu_{x_m \to f_s}(x_m) = \sum_{X_{sm}} [\prod_{l \in ne(x_m)\backslash f_s} F_l(x_m, X_{ml})] = \prod_{l \in ne(x_m)\backslash f_s} [\sum_{X_{ml}} F_l(x_m, X_{ml}$

- $$p(X_s) = f_s(x_s) \prod_{i \in ne(f_s)} \mu_{x_i \to f_s}(x_i)$$
- If aim to obtain the marginal distribution of all variables in the factor graph, we need not to run this algorithm for |Variable| times, just like the following figure:



(a)

(b)

TrueSkill[13][14]

Bayesian Linear Probit Model & Derivations[15][16]

- A specific situation of TrueSkill
- $P(y \mid x, \tilde{\mu}) = \Phi\left(y \cdot \frac{\tilde{\mu}^T x}{\beta}\right)$, where $\Phi(t) := \int_{-\infty}^{t} \mathcal{N}(s; 0, 1) ds$
- Notations:
  1. A vector of means $\mu := (\mu_{1,1}, \ldots\ldots, \mu_{N,M_N})^T$
  2. A vector of variances $\sigma^2 := (\sigma_{1,1}^2, \ldots\ldots, \sigma_{N,M_N}^2)^T$
  3. sample $x := (x_1^T, \ldots\ldots, x_N^T)$,
     $x_i := (x_{i,1}, \ldots\ldots, x_{i,M_i}), \sum_{j=1}^{M_i} x_{i,j} = 1$
  4. $\sum^2 := \beta^2 + x^T \sigma^2$
  5. $v(t) := \frac{\mathcal{N}(t;0,1)}{\Phi(t;0,1)}$ $\quad w(t) := v(t)[v(t) + t]$
- Update Formulation:

$$\tilde{\mu}_{i,j} = \mu_{i,j} + y x_{i,j} \frac{\sigma_{i,j}^2}{\Sigma} v\left(\frac{y x^T \mu}{\Sigma}\right)$$

$$\tilde{\sigma}_{i,j}^2 = \sigma_{i,j}^2 \left(1 - x_{i,j} \frac{\sigma_{i,j}^2}{\Sigma^2} w\left(\frac{y x^T \mu}{\Sigma}\right)\right)$$

Bayesian Meta-Prior Learning Using Empirical Bayes

Starting Point

1. Search for a good design of informative prior instead a non-informative prior
2. decouple the learning rate of different categories
3. Solution to the trade-off between exploration and exploitation
4. Provide a general solution in estimating empirical priors for a wide range of applications that are modeled as Bayesian bandits or involve Bayesian learning, esp. recommender system and MABs

Objective

Utilize early randomized data and Empirical Bayes(EB) method to construct an experiment-specific hierarchical informative prior

Empirical Prior Derivation and Estimation

- Problem of interest: features could be grouped into arbitraty non-overlapping categories
- Assumptions:
    1. The kth category C_k has a distince hyperparameter meta-prior distribution(determined by experts' knowledge), here we assume it as Gaussian, $N(\nu_k, \tau_k^2)$
    2. One feature's true effect(i.e. coefficients) μ_i is drawn i.i.d from the corresponding category's meta-prior distribution: $\mu_i \sim N(\nu_k, \tau_k^2)$ and $\mathbb{E}[\mu_i] = \nu_k$,$\mathbb{V}[\mu_i] = \tau_k^2 \ \forall i \in C_k$
    3. $\tilde{\mu_i}$ and $\widetilde{\sigma_i^2}$ is the estimators of $\mathbb{E}[\mu_i]$ and $\mathbb{V}[\mu_i]$ respectively.
    4. $\mathbb{E}\left[\tilde{\mu}_i \mid \mu_i\right] = \mu_i, \mathbb{V}\left[\tilde{\mu}_i \mid \mu_i\right] = \tilde{\sigma}_i^2, \mathbb{E}\left[\tilde{\mu}_i\right] = \nu_k, \quad \forall i \in C_k$

5. Setting $\nu_k = 0$ to ensure the model is invariant to input feature sign changes.
6. $y \in \{-1, 1\}$ denotes the response binary variable, x denotes the features

- Derivation and Estimation of hyperparameters
    1. Variance Decomposition: $\mathbb{V}\left[\tilde{\mu}_i\right] = \mathbb{E}\left[\mathbb{V}\left[\tilde{\mu}_i \mid \mu_i\right]\right] + \mathbb{V}\left[\mathbb{E}\left[\tilde{\mu}_i \mid \mu_i\right]\right] = \mathbb{E}\left[\tilde{\sigma}_i^2\right] + \tau_k^2, \quad \forall i \in C_k$
    2. $\tau_k^2 = \mathbb{V}\left[\tilde{\mu}_i\right] - \mathbb{E}\left[\tilde{\sigma}_i^2\right]$
    3. $\hat{\tau}_{k,t}^2 = \widehat{\mathbb{V}\left[\tilde{\mu}_{i,t}\right]} - \widehat{\mathbb{E}\left[\tilde{\sigma}_{i,t}^2\right]} = \frac{\sum_{i \in C_k} \left(\tilde{\mu}_{i,t} - \hat{\nu}_{k,t}\right)^2}{N_k - 1} - \frac{\sum_{i \in C_k} \tilde{\sigma}_{i,t}^2}{N_k}$ as an estimator for the meta-prior variance at time t
    4. $\hat{\nu}_{k,t} = \frac{\sum_{i \in C_k} \tilde{\mu}_{i,t}}{N_k}, \quad \forall C_k$
    5. Setting $\nu_k = 0$ and obtain additional one degree of freedom
    6. $\hat{\tau}_{k,t}^2 = \frac{\sum_{i \in C_k} \left[\tilde{\mu}_{i,t}^2 - \tilde{\sigma}_{i,t}^2\right]}{N_k}, \quad \forall C_k$
- Properties of Empirical Prior
    1. Unbiasedness
    2. Strong consistency
    3. Bandit Cumulative Upper Bound

Models

Meta-Prior + Bayesian Linear Probit Regression: $P(y \mid x, \tilde{\mu}) = \Phi\left(y \cdot \frac{\tilde{\mu}^T x}{\beta}\right) \quad \beta = 1$

Experiments

Adult Dataset

- One-hot encoding for first and second order features
- Variable selection through adaptive LASSO and two choice for selections:

    1. select only the variables with non-zero weights
    2. select the whole feature when there exists a non-zero weight interaction in it
- Problems about degenerative meta-prior variance
    1. Speculation of low traffic of a training batch
    2. Some solutions toward this: Bootstrapping, Ensembles, Epoch Training
    3. Bootstrap for the first batch into several 5K instances batches, treat each bootstrapped set as a training epoch and compute the meta-prior variance through standard Gaussian prior after each epoch until obtaining the non-degenerative meta-prior variance
- Three Models:
    1. BLIP: Update the model in batch with day t data
    2. BLIPBayes: Specify the prior reset time t and upgrade the meta-prior variance using the bootstrapped data from the data observed until t
    3. BLIPTwice: update the model twice, first with the same bootstrapped data as BLIPBayes and second with all the data observed until day t
       Choose cross-entropy as criterion to compute the log loss of the testing set
- Simulation Results
    1. Problems:
        - The final results derive the opposite conclusion of the article
        - Encounter the degenerative variance of prior when updating the parameters

MAB Live Experiments

Conclusions and Individual Method

Conclusions

- We've understood the method and details in the article "Bayesian Meta-Prior Learning Using Empirical Bayes", including empirical bayes, multi-arm bandits in recommender systems which focuses on the Exploration-Exploitation dilemma and reflects the method of reinforcement learning, meta-learning which manifests the relationship between the optimization and bayesian hierarchical model, bayesian linear probit model derived via the mechanism of factor graph and sum-product algorithm, where TrueSkill is one of its famous applications.

Individual Method

---

1. Empirical Bayes: Concepts and Methods, https://efron.ckirby.su.domains/papers/2021EB-concepts-methods.pdf ↩
2. https://arxiv.org/abs/0908.1709 ↩
3. https://arxiv.org/abs/0908.1712 ↩
4. http://www3.stat.sinica.edu.tw/sstest/oldpdf/A22n313.pdf ↩
5. https://arxiv.org/abs/1712.02009 ↩
6. Empirical Bayes mean estimation with nonparametric errors via order statistic regression on replicated data ↩
7. https://arxiv.org/abs/1902.02774 ↩
8. RECASTING GRADIENT-BASED META-LEARNING AS HIERARCHICAL BAYES, https://openreview.net/pdf?id=BJ_UL-k0b ↩
9. Gaussian Imagination in Bandit Learning, https://arxiv.org/abs/2201.01902 ↩

10. Expectation Propagation for Bayesian Inference, https://danielkhashabi.com/learn/ep.pdf ↩
11. Pattern Recognition and Machine Learning, Chapter 8 ↩
12. Factor Graphs & Sum Product Algorithm https://www.cs.toronto.edu/~radford/csc2506/factor.pdf ↩
13. TrueSkill: A Bayesian Skill Rating System, https://proceedings.neurips.cc/paper/2006/file/f44ee263952e65b3610b8ba51229d1f9-Paper.pdf ↩
14. The Math Behind TrueSkill, https://www.moserware.com/assets/computing-your-skill/The%20Math%20Behind%20TrueSkill.pdf ↩
15. Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine, https://quinonero.net/Publications/AdPredictorICML2010-final.pdf ↩
16. Practical Lessons from Predicting Clicks on Ads at Facebook, https://quinonero.net/Publications/predicting-clicks-facebook.pdf ↩