

## 1 Maximum Likelihood Estimator

- Consistent, asymptotic unbiased.  $\hat{\theta}_n^{\text{MLE}} \xrightarrow{P} \theta_0$ .
- Asymptotic normal.  $\sqrt{n}(\hat{\theta}_n^{\text{MLE}} - \theta_0) \xrightarrow{D} \mathcal{N}(0, I_n)$ ,  $I_n(\theta_0) = \mathbb{E}[-\ddot{s}_\theta] = \mathbb{E}[s_\theta s_\theta^\top]$  Fisher info,  $s_\theta(x) = \partial_\theta \ln p_\theta(x)$  score func,  $\mathbb{E}_\theta s_\theta = 0$ .
- Asymptotic efficient.  $\hat{\theta}_n^{\text{MLE}}$  reaches CRLB when  $n \rightarrow \infty$ .  $\hat{\theta}_n^{\text{MLE}}$  not efficient if  $n$  finite. Stein estimator always better.
- Equivariance, if  $\hat{\theta}_n$  is MLE,  $\hat{y} = g(\hat{\theta}_n^{\text{MLE}})$  is MLE of  $\mathcal{L}(g^{-1}(y))$ . Proof by optimality of MLE. Cramer-Rao lower bound (CRLB): for any unbiased  $\hat{\theta}$  of  $\theta_0$ ,  $\mathbb{E}(\hat{\theta} - \theta_0)^2 \geq 1/I_n(\theta_0)$   
Proof:  $\text{Cov}[s_\theta, \hat{\theta}] = \mathbb{E}[s_\theta, \hat{\theta}] = \partial_\theta \mathbb{E}[\hat{\theta}] = \partial_\theta \theta = 1$ . Cauchy-Schwarz  $\text{Cov}^2[s_\theta, \hat{\theta}] \leq \text{Var}[s_\theta] \text{Var}[\hat{\theta}] = I_n(\theta) \mathbb{E}(\hat{\theta} - \theta_0)^2$  QED.  
However, when dimension of problem goes to infinity while data-dim ratio is fixed, MLE is biased and the  $p$ -values are unreliable.

## 2 Regression

### Bias-Variance trade-off

Training dataset,  $\hat{f}$  be the predictive function.  $\mathbb{E}_D \mathbb{E}_{Y|X} (\hat{f}(X) - Y)^2 = \mathbb{E}_D (\hat{f}(x) - \mathbb{E}_D \hat{f}(x))^2 + (\mathbb{E}_D \hat{f}(x) - \mathbb{E}(Y|X))^2 + \mathbb{E}_D (\mathbb{E}(Y|X) - Y)^2 = \text{Model Variance} + \text{Bias}^2 + \text{Intrinsic Noise}$ . The optimal trade-off is achieved by avoiding under-fitting (large bias) and over-fitting (large variance). Note that here the variance of output is computed by refitting the regressor on a new dataset.

### Regularization

Ridge and Lasso can be viewed as MAP estimation with a prior on  $\beta$ . Ridge has a Gaussian Prior and LASSO a Laplacian prior. Using SVD, we get Ridge has built-in model selection:  $X\beta^{\text{Ridge}} = \sum_{j=1}^d [d_j^2/(d_j^2 + \lambda)] u_j u_j^\top Y$  (each  $u_j u_j^\top Y$  can be viewed as a model). Lasso has more sparse estimations because the gradient of regularization does not shrink as Ridge.

## 3 BLR and GP

### Bayesian Linear Regression

$Y = X\beta + \epsilon \sim \mathcal{N}(0, \sigma^2)$ . Prior  $\beta \sim \mathcal{N}(0, \Lambda^{-1})$ . Posterior  $\beta|X, Y \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$ ,  $\Sigma_\beta = (\sigma^{-2} X^\top X + \Lambda)^{-1}$ ,  $\mu_\beta = \sigma^2 \Sigma_\beta X^\top Y$ .

### Gaussian Process

$Y = \begin{pmatrix} Y_0 \\ Y_1 \end{pmatrix}$  is the combination of observed and prediction value. Assume a Gaussian prior of  $\mathcal{N}(0, K + \sigma^2 I)$ , where  $K_{ij} = k(x_i, x_j)$  is kernel. GP regression is the conditional/Posterior distribution on  $Y_0$ ,  $\mathbb{E}[Y_1|Y_0] = K_{10}(\sigma^2 I_0 +$

$K_{00})^{-1} Y_0$ ,  $\text{Cov}[Y_1] = \sigma^2 I_1 + K_{11} - K_{10}(\sigma^2 I_0 + K_{00})^{-1} K_{01}$ . Bayesian LR is a special case of GP with linear kernel  $k(x, y) = x^\top \Lambda^{-1} y$ .

### Kernel Function

A function is a kernel iff (1) symmetry  $k(x, x') = k(x', x)$  and (2) semi-positive definite  $\int_{\Omega} k(x, x') f(x) f(x') dx dx' \geq 0$  for any  $f \in L_2$  and  $\Omega \in \mathcal{R}^d$  (continuous) or  $K(X) \geq 0$  (discrete). The latter is equivalent to (1)  $a^\top K a \geq 0, \forall a$  or (2)  $k(x, x') = \phi(x)^\top \phi(x')$  for some  $\phi$ .

### Kernel Construction

If  $k_{1,2}$  are valid kernels, then followings are valid: (1)  $k(x, x') = k_1(x, x') + k_2(x, x')$ . (2)  $k(x, x') = k_1(x, x') \cdot k_2(x, x')$ . Proof: expand by Mercer's thm. (3)  $k(x, x') = c k_1(x, x')$  for constant  $c > 0$ . (4)  $k(x, x') = f(k_1(x, x'))$  if  $f$  is a polynomial with positive coefficients or the exp. Proof: polynomial can be proved by applying the product, positive scaling and addition. Exp can be proved by taking limit on the polynomial. (5)  $k(x, x') = f(x) k_1(x, x') f(x')$ . (6)  $k(x, x') = k_1(\phi(x), \phi(x'))$  for any function  $\phi$ .

Example: RBF kernel  $k(x, y) = e^{-\|x-y\|^2/2\sigma^2} = e^{-\|x\|^2/2\sigma^2} \times e^{x^\top y/2\sigma^2} \times e^{-\|y\|^2/2\sigma^2}$  is valid. (1)  $x^\top y$  linear kernel is valid (2) then  $\exp(\frac{1}{\sigma^2} x^\top y)$  is valid, (3) let  $f(x) = \exp(-\frac{1}{2\sigma^2} \|x\|^2)$ , by rules  $f(x)k(x, y)f(y)$  RBF is valid.

**Mercer's Theorem:** Assume  $k(x, x')$  is a valid kernel. Then there exists an orthogonal basis  $e_i$  and  $\lambda_i \geq 0$ , s.t.  $k(x, x') = \sum_i \lambda_i e_i(x) e_i(x')$ .

## 4 Linear Methods for Classification

### Concept Comparison

- Probabilistic Generative, modeling  $p(x, y)$ : (1) can create new samples, (2) outlier detection, (3) probability for prediction, (4) high computational cost and (5) high bias.
- Probabilistic Discriminative, modeling  $p(y|x)$ : (1) probability for prediction, (2) medium computational cost and (3) medium bias.
- Discriminative, modeling  $y = f(x)$ : (1) no probability for prediction, (2) low computational cost and (3) low bias.

### Infer $p(x, y)$ for classification problems

Use  $p(x, y) = p(y)p(x|y)$ . Since  $y$  has finite states, model  $p(y)$  and  $p(x|y)$  for different  $y$ . The modeling requires to (1) guess a distribution family and (2) infer param by MLE.

### Compute $p(y|x)$ by discriminant analysis (DA)

**Linear DA** Assumption: classify a sample into two Gaussian distribution with  $\Sigma_0 =$

$\Sigma_1$ . After calculation,  $p(y = 1|x) = 1/(1 + \exp(-\log \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)})) = 1/(1 + \exp(w_1^\top x + w_0))$  since the quadratic term is eliminated due to  $\Sigma_0 = \Sigma_1$ .

**Quadratic DA** Assumption: classify a sample into two Gaussian distribution with  $\Sigma_0 \neq \Sigma_1$ . After calculation,  $p(y = 1|x) = 1/(1 + \exp(x^\top W x + w_1^\top x + w_0))$ .

### Optimization Methods

*Optimal Learning Rate for Gradient Descent*

Goal: find  $\eta^* = \argmin_{\eta} L(w^k - \eta \cdot \nabla L(w^k))$ .

By Taylor expansion of  $L(w^{k+1})$  at  $w^k$  and solve for the optimal  $\eta$ , we get  $\eta^* = \|\nabla L(w^k)\|^2 / (\nabla L(w^k)^\top H_L(w^k) \nabla L(w^k))$ .

Cons of naive gradient descent: (1) zig-zag behavior, especially in a very narrow, long and slightly downward valley; (2) gradient update is small near the stationary point. Mitigated by adding momentum into update,  $w^{k+1} = w^k - \eta \nabla L(w^k) + \mu^k (w^k - w^{k-1})$ , this speeds update towards "common" direction.

### Newton's Method

Taylor-expand  $L(w)$  at  $w_k$  to derive the optimal  $w^{k+1}$ :  $L(w) \approx L(w^k) + (w - w^k)^\top \nabla L(w^k) + \frac{1}{2} (w - w^k)^\top H_L(w^k) (w - w^k) \Rightarrow w^{k+1} = w^k - H_L^{-1}(w^k) \nabla L(w^k)$ .

Pros: (1) better updates compared to GD since it uses the second Taylor term and (2) does not require learning rate.

Cons: requires  $H_L^{-1}$  which is expensive.

### Bayesian Method

In most cases, the posterior is intractable. Use approximation of posterior instead.

### Laplacian Method

Idea: approximate posterior near the MAP estimation with a Gaussian distribution.  $p(w|X, Y) \propto p(w, X, Y) \propto \exp(-R(w))$ , where  $R(w) = -\log p(w, X, Y)$ . Let  $w^* = \argmin R(w)$  be the MAP estimation and Taylor-expand  $R(w)$  at  $w^*$ :  $R(w) \approx R(w^*) + \frac{1}{2} (w - w^*)^\top H_R(w^*) (w - w^*)$ . Therefore,  $p(w|X, Y) \propto \exp(-R(w^*) - \frac{1}{2} (w - w^*)^\top H_R(w^*) (w - w^*))$  and thus  $(w|X, Y) \sim \mathcal{N}(w^*, H_R^{-1}(w^*))$ .

### AIC & BIC

- Define  $\text{BIC} = k \log N - 2 \log \hat{L}$ , where  $k$  is #parameters and  $\hat{L}$  is the likelihood  $p(x|w^*)$ . A lower BIC means a better model.
- Define  $\text{AIC} = 2k - 2 \log \hat{L}$ . A lower AIC means a better model.

## LDA by loss minimization

**Perceptron** for  $y_i \in \{0, 1\}$ , find  $w$ , s.t.  $y_i w^\top x_i > 0$  for any  $i$ . Prediction is  $c(x) = \text{sgn}(w^\top x)$ .

$L(y, c(x)) = \min\{0, -y w^\top x\}$ . By GD  $w^{(k+1)} \leftarrow w^{(k)} + \eta(k) \sum_i \text{wrong } y_i x_i$ , Perceptron will converge if (1) data linearly separable, (2) learning rate  $\eta(k) > 0$ , (3)  $\sum_k \eta(k) \rightarrow +\infty$  and (4)  $(\sum_k \eta(k)^2) / (\sum_k \eta(k))^2 \rightarrow 0$ . However, multiple solutions permitted if data linearly separable, solution unstable.

### Fisher's LDA

Idea: project the two distribution into one dimension and maximize the ratio of the variance between the classes and the variance within the classes, i.e.,  $\max(w^\top u_1 - w^\top u_0)^2 / (w^\top S w)$ , where  $S = \Sigma_0 + \Sigma_1$ . Let gradient be zero and solve for  $w^*$ , we get  $w^* \propto S^{-1}(u_1 - u_0)$ .

We first compute  $w^*$  and fit distributions of the two-class projection. Then apply Bayesian decision theory to make classification.

## 5 Optimization with Constraint

**Problem**  $\min_x f(x)$  s.t.  $g_i \in [I](x) \leq 0$  and  $h_j \in [J](x) = 0$ . Solve it with **KKT Cond**: (1) Stationary  $\nabla f + \sum_i \lambda_i \nabla g_i + \sum_j \mu_j \nabla h_j = 0$ , (2)  $h_j(x) = 0$ , (3) primal feasibility  $g_i(x) \leq 0$ , (4) dual feasibility  $\lambda_i \geq 0$ , (5) complementary slackness  $\lambda_i g_i(x) = 0$ .

**Weak Duality:** Lagrangian  $L(x, \lambda, \mu) = f(x) + \lambda^\top g(x) + \mu^\top h(x)$ ,  $\lambda > 0$ . Dual function  $F(\lambda, \mu) := \min_x L(x, \lambda, \mu)$ . Denote  $\tilde{x}$  optima of original problem, then  $\lambda^\top g(\tilde{x}) + \mu^\top h(\tilde{x}) \leq 0, \forall \lambda, \mu$ ,  $F(\lambda, \mu) = \min_x L(x, \lambda, \mu) \leq L(\tilde{x}, \lambda, \mu) \leq f(\tilde{x}) = \min_{x, h(x)=0, g(x) \leq 0} f(x)$

### Strong Duality in Convex Optimization

If **Slater's cond** (1)  $f$  convex (2)  $g$  convex (3)  $h$  linear (4)  $\exists \bar{x}$  s.t.  $g_i(\bar{x}) < 0$  and  $h_j(\bar{x}) = 0$ , then Strong Duality  $\max_{\lambda, \mu} F(\lambda, \mu) = \min_{x, h(x)=0, g(x) \leq 0} f(x)$  holds.

## 6 Support Vector Machine

### Linear Separable Case

**Primal:**  $\max_{w, b, t} t$  s.t.  $\forall i, t \leq y_i (w^\top x_i + b)$  and  $\|w\| = 1$   
 $\Leftrightarrow \min_{w, b} \frac{1}{2} w^2$  s.t.  $\forall i, 1 \leq y_i (w^\top x_i + b)$   
(1) **KKT cond:**  $\forall i, \alpha_i \geq 0, (1 - y_i (w^\top x_i + b)) \leq 0, \alpha_i (1 - y_i (w^\top x_i + b)) = 0$   
(2) **Dual:**  $\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$  s.t.  $(\alpha_i \geq 0) \wedge (\sum_i \alpha_i y_i = 0)$

## Non-separable Case

Introduce slack variables  $\xi_i := \max\{1 - y_i(w^\top x_i + b), 0\} = [1 - y_i(w^\top x_i + b)]_+$  into loss.

**Primal:**  $\min_{w,b} \frac{1}{2}w^2 + C \sum_i \xi_i = \min_{w,b} \frac{1}{2}w^2 + C[1 - y_i(w^\top x_i + b)]_+$ . Hinge loss  $[1 - x]_+$ .

Equivalent form:  $\min_{w,b} \frac{1}{2}w^2 + C \sum_i \xi_i$  s.t.  $y_i(w^\top x_i + b) \geq 1 - \xi_i$  and  $\xi_i \geq 0$

**Dual:**  $\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$  s.t.  $\sum_i \alpha_i y_i = 0$  and  $0 \leq \alpha_i \leq C$

## Multi-class SVM

$\min_{w=[w_{0:K-1}], b=[b_{0:K-1}]} \frac{1}{2}\|w\|^2 + \sum_i C \xi_i$  s.t.  $\xi_i \geq 0$  and  $(w_y^\top x + b_y) - (w_{y'}^\top x + b_{y'}) \geq 1 - \xi_i, \forall y \neq y_i$

## Structural SVM

$y$  is structured, e.g. trees, maximum margin between  $y_i, y_j$  depends on their similarity, so the condition changes to  $w^\top \Psi(x_i, y_i) - w^\top \Psi(x_i, y) \geq \Delta(y_i, y) - \xi_i, \forall y \neq y_i$ .

## 7 Ensemble

**Bagging** Each bagged estimator have bias  $\beta = \mathbb{E}(y - b(x))^2$ , variance  $\sigma^2 = \text{Var}b(x)$  co-variance  $\rho^2 = \text{Cov}(b(x), b'(x))/\sigma^2$ . Then  $\mathbb{E}(y - \sum_m b^{(m)}(x)/M)^2 = \beta^2 + \sum_m \mathbb{E}(\beta - b^{(m)}(x))^2/M^2 = \beta^2 + \sigma^2/M + \sigma^2 \rho^2(1 - 1/M)$ . In class we assume  $\rho = 0$ . Anyway Bagging reduces variance. Random Forest is a case of Bagging. Bagging induces implicit regularization.

**Adaboost** Initial  $w_i^{(0)} = 1/n$ . For  $t \in [M]$ , (1) train  $f_t(x) = \text{argmin}_{b(x)} \sum_i w_i^{(t)} \mathbb{I}_{\{y_i \neq b(x_i)\}}$  (2) error  $\epsilon_t = (\sum_i w_i^{(t)} \mathbb{I}_{\{y_i \neq f_t(x_i)\}}) / \sum_i w_i^{(t)}$  (3) estimator weight  $\alpha_t = \log(\frac{1-\epsilon_t}{\epsilon_t})$  (4) data weight  $w_i^{(t+1)} = w_i^{(t)} e^{\alpha_t \mathbb{I}_{\{y_i \neq f_t(x_i)\}}}$

**Prediction**  $\hat{c} = \text{sgn}(\sum_{t=1}^M \alpha_t f_t(x))$

**Gradient Boosting** Initial  $f_0(x) = 0$ . For  $t \in [M]$ , (1) train  $(\alpha_t, b^{(t)}) \leftarrow \text{argmin}_{\alpha > 0, b \in \mathcal{H}} \sum_{i=1}^n L(y_i, \alpha b(x_i) + f_{t-1}(x_i))$  (2) update function  $f_t(x) \leftarrow \alpha_t b^{(t)}(x) + f_{t-1}(x)$ . **Prediction**  $\hat{c}(x) = \text{sgn}(f_M(x))$ . Adaboost is GB with  $L(y, \hat{y}) = e^{-y\hat{y}}$ .

## 8 Generative Models

**ELBO**  $\ln p(y) = \ln \int p(y | \theta) p(\theta) d\theta = \ln \mathbb{E}_{\theta \sim q} \left[ p(y | \theta) \frac{p(\theta)}{q(\theta)} \right] \geq \mathbb{E}_{\theta \sim q} \left[ \ln \left( p(y | \theta) \frac{p(\theta)}{q(\theta)} \right) \right] = \mathbb{E}_{\theta \sim q} [\ln p(y | \theta)] - KL(q || p(\cdot))$

**VAE Goal:** Find a latent representation  $z$  of  $x$  with simple prior  $p_\theta(z)$ . Problem:  $p_\theta(x) = \mathbb{E}_\theta p(x|z)$  intractable. Solution: use encoder net  $q_e(x|z)$  and  $q_d(z|x)$  to model conditional and posterior prob.

**ELBO for VAE training** loss  $l = \sum \ln(p_\theta(x_i))$

$$\begin{aligned} \ln(p_\theta(x_i)) &= \mathbb{E}_{z \sim q_\phi(z|x_i)} [\ln p_\theta(x_i)] = \mathbb{E}_Z \left[ \ln \frac{p_\theta(x_i | z) p_\theta(z)}{p_\theta(z | x_i)} \right] \\ &= \mathbb{E}_Z \left[ \ln \frac{p_\theta(x_i | z) p_\theta(z)}{p_\theta(z | x_i)} \frac{q_\phi(z | x_i)}{q_\phi(z | x_i)} \right] \\ &= \mathbb{E}_Z [\ln p_\theta(x_i | z)] - \mathbb{E}_Z \left[ \ln \frac{q_\phi(z | x_i)}{p_\theta(z)} \right] + \mathbb{E}_Z \left[ \ln \frac{q_\phi(z | x_i)}{p_\theta(z | x_i)} \right] \\ &= \underbrace{\mathbb{E}_Z [\ln p_\theta(x_i | z)] - \text{KL}(q_\phi(z | x_i) || p_\theta(z))}_{\text{ELBO } \mathcal{L}(x_i, \theta, \phi)} \end{aligned}$$

+  $\text{KL}(q_\phi(z | x_i) || p_\theta(z | x_i)) \geq \text{ELBO}$ .

**Generative Adversarial Network:** Generator  $G$  and Discriminator  $D$ . Optimize  $\min_G \max_D V(D, G)$  where  $V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\ln D(x)] + \mathbb{E}_{z \sim p_z(z)} [\ln(1 - D(G(z)))]$

## 9 Convergence of SGD, Robbins-Monro

Loss gradient  $\ell(\cdot)$ , SGD update  $z^{(t)} \leftarrow \ell(\theta^{(t)}) + \gamma^{(t)}, \theta^{(t+1)} \leftarrow \theta^{(t)} - \eta^{(t)} z^{(t)}, \gamma^{(t)}$  noise.

Problem: Whether  $\theta^\infty \rightarrow \arg_{\theta^*} \mathbb{E}[\ell(\theta^*)] \triangleq 0$ ?

Assume: (1)  $\mathbb{E}[\gamma] = 0$ , (2)  $\mathbb{E}[\gamma^2] = \sigma$  (3)  $(\theta - \theta^*)\ell(\theta) > 0, \forall \theta \neq \theta^*$  (4)  $\exists b, \ell(\theta) < b, \forall \theta$ . If (1)  $\eta^{(t)} \rightarrow 0$  (2)  $\sum \eta^{(t)} = \infty$  (3)  $\sum \eta^{(t)^2} < \infty$ ,

then  $\mathbb{P}(\theta^* = \theta^{(t)}) \xrightarrow{t \rightarrow \infty} 1$ .

Proof:  $\mathbb{E}[(\theta^{(t+1)} - \theta^*)^2] = \mathbb{E}[(\theta^{(t)} - \theta^*) - \eta^{(t)} \ell(\theta^{(t)}) - \eta^{(t)} \gamma^{(t)}]^2$ .  $\gamma^{(t)}$  independent with  $\theta^{(t)}, \ell(\theta^{(t)})$ , so LHS =  $\mathbb{E}[(\theta^* - \theta^{(t)})^2] - 2\eta^{(t)} \mathbb{E}[\ell(\theta^{(t)})](\theta^* - \theta^{(t)}) + \eta^{(t)^2} (\mathbb{E}[\ell^2(\theta^{(t)})] + \mathbb{E}[\gamma^2(t)]) \leq \mathbb{E}[(\theta^* - \theta^{(0)})^2] - 2 \sum_{i \leq t} \eta^{(i)} \mathbb{E}[\ell(\theta^{(i)})](\theta^* - \theta^{(i)}) + \sum_{i \leq t} \eta^{(i)^2} (b^2 + \sigma^2)$  Since  $0 \leq \mathbb{E}[(\theta^* - \theta^{(t+1)})^2] < \infty, 0 = \lim_{i \rightarrow \infty} \mathbb{E}[\ell(\theta^{(i)})](\theta^* - \theta^{(i)}) = \lim_{i \rightarrow \infty} \mathbb{P}(\theta^* = \theta^{(i)}) \mathbb{E}[\ell(\theta^{(i)})](\theta^* - \theta^{(i)}) | \theta^* = \theta^{(i)} + \mathbb{P}(\theta^* \neq \theta^{(i)}) \mathbb{E}[\ell(\theta^{(i)})](\theta^* - \theta^{(i)}) | \theta^* \neq \theta^{(i)}, \lim_{i \rightarrow \infty} \mathbb{P}(\theta^* \neq \theta^{(i)}) = 0$

## 10 Non-parametric Bayesian Inference (BI) Exact Conjugate Prior of Multivariate Gaussian

Data:  $x_i \sim \mathcal{N}(\mu, \Sigma)$  i.i.d.. Inverse Wishart:  $\Sigma \sim \mathcal{W}^{-1}(S, \nu) \propto |\Sigma|^{(\nu+p+1)/2} \exp(-\text{Tr}(\Sigma^{-1}S)/2)$ .

**Normal Inverse Wishart** as conjugate prior:  $p(\mu, \Sigma | m_0, k_0, \nu_0, S_0) = \mathcal{N}(\mu | m, \Sigma/k_0) \mathcal{W}^{-1}(\Sigma | S_0, \nu_0)$  Update rule:  $m_p = (k_0 m_0 + N \bar{x}) / (k_0 + N)$ ,  $k_p = k_0 + N$ ,  $\nu_p = \nu_0 + N$ ,  $S_p = S_0 + k_0 m_0 m_0^\top - k_p m_p m_p^\top + \sum (x_i - \bar{x})(x_i - \bar{x})^\top$ .

## BI with Semi-Conjugate Prior

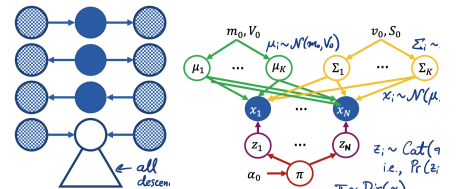
New prior:  $\mu \sim \mathcal{N}(m_0, V_0)$ ,  $\Sigma \sim \mathcal{W}^{-1}(S_0, \nu_0)$ , then posterior  $p(\mu, \Sigma | X)$  is intractable, but condition posterior is exact,  $p(\mu | \Sigma, X) = \mathcal{N}(m_p, V_p)$ ,  $V_p^{-1} = V_0^{-1} + N \Sigma^{-1}$ ,  $V_p^{-1} m_p = V_0^{-1} m_0 + N \Sigma^{-1} \bar{x}$ ;  $p(\Sigma | \mu, X) = \mathcal{W}^{-1}(S_p, \nu_p)$ ,  $\nu_p = \nu_0 + N$ ,  $S_p = S_0 + \sum x_i x_i^\top + N \mu \mu^\top - 2N \bar{x} \mu^\top$ .

**Gibbs sampling:** random variable  $p(z_1, \dots, z_p)$  intractable, cyclically resample  $z_i$  according to tractable conditional distribution  $p(z_i | z_{\setminus i})$   $n$  times, when  $n \rightarrow \infty$ ,  $(z_1, \dots, z_p) \sim p(z_1, \dots, z_p)$  Finally, replace posterior with MC sampling:  $\mathbb{E}_{\theta | X} f(x | \theta) \approx \sum f(x | \theta_i) / N$

## BI for Gaussian Mixture Model

Data model: latent  $K$  class variable  $z_i \sim \text{Cat}(\pi)$ , observed  $x_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$ . Prior:  $\mu_k \sim \mathcal{N}(m_0, V_0)$ ,  $\Sigma_k \sim \mathcal{W}^{-1}(S_0, \nu_0)$ ,  $\pi \sim \text{Dir}(\alpha) \propto \prod_k p_k^{\alpha_k - 1}$ . Prior also intractable.

**Goal** Gibbs sampling for BI, but to simplify conditional distribution.



**d-seperation:** for verifying conditional independence. Given with observed variable set  $C$ , if every path from variable  $A$  to  $B$  is blocked on probability graph, then  $A$  and  $B$  are independent condition on  $C$ . By this thm: (1)  $z_i, z_j$  (2)  $\mu, \pi$  (3)  $\Sigma, \pi$  all independent condition on other parameter. Sampling procedure: (1)  $z^{(t)} \leftarrow p(\cdot | x, \mu^{(t-1)}, \Sigma^{(t-1)})$ , (2)  $\mu^{(t)} \leftarrow p(\cdot | x, \Sigma^{(t-1)}, z^{(t)})$ , (3)  $\Sigma^{(t)} \leftarrow p(\cdot | x, \mu^{(t)}, z^{(t)})$ , (4)  $\pi^{(t)} \leftarrow p(\cdot | x, z^{(t)})$

## BI for Non-Parametric GMM

**Goal:** sample from infinite categorical distri. **Dirichlet Process (DP):**  $\Theta$  parameter space,  $H$  prior distri on  $\Theta$ ,  $A_1, \dots, A_r$  arbitrary partition of  $\Theta$ .  $G$  a categorical distribution over  $\{A_i\}$  is  $G \sim \text{DP}(\alpha H)$  if  $(G(A_1), \dots, G(A_r)) \sim \text{Dir}(\alpha H(A_1), \dots, \alpha H(A_r))$ .

**Posterior:**  $G | \{\theta_i\}_{i=1}^n \sim \text{DP}\left(\alpha + n, \frac{\alpha H + \sum_{i=1}^n \delta_{\theta_i}}{\alpha + n}\right)$

**Condition on  $\theta$ , Margin over  $G$ :**  $\theta_{n+1} | \theta_1, \dots, \theta_n \sim \frac{1}{\alpha + n} (\alpha H + \sum_{i=1}^n \delta_{\theta_i})$ , Leads to CRP

## Three Methods of Sampling from DP

In  $K \rightarrow \infty$  GMM,  $\theta$  in DP is  $z$ ,  $G$  is  $\pi$ .

**(1) Chinese Restaurant Process (CRP)**, sample  $z$ , marginalize over  $\pi$ :

$$p(z_n = k | \theta_{i < n}) = \begin{cases} n_k / (\alpha + n - 1), & \text{existing } k \\ \alpha / (\alpha + n - 1), & \text{new } k \end{cases}$$

**Expect # of Class**  $\sum_{i=1}^n \frac{\alpha}{\alpha + i - 1} \approx e q \alpha \log\left(1 + \frac{n}{\alpha}\right)$

**(2) Stick-breaking Construction** samples  $\pi$ :

$$\beta_k \sim \text{Beta}(1, \alpha), \theta_k^* \sim H, \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l)$$

**(3) Marginalize over  $\mu, \Sigma$**  when sampling  $z$  (if intractable), less variance (Rao-Blackwall).

**Exchangeability:**  $p(\{\theta_i\}) = \prod_{n=1}^N p(\theta_n | \{\theta_{i < n}\})$  unchanged after permuting sampling order.

**DeFinetti's Thm** any exchangeable distri is a mixture model  $P(\{\theta_i\}) = \int \prod_{i=1}^n G(\theta_i) dP(G)$

## 11 PAC Learning

• A learning algorithm  $\mathcal{A}$  can learn  $c \in \mathcal{C}$  if there is a poly( $\dots$ ), s.t. for (1) any distribution  $\mathcal{D}$  on  $\mathcal{X}$  and (2)  $\forall \epsilon \in [0, 1/2], \delta \in [0, 1/2]$ ,  $\mathcal{A}$  outputs  $\hat{c} \in \mathcal{H}$  given a sample of size at least  $\text{poly}(\frac{1}{\epsilon}, \frac{1}{\delta}, \text{size}(c))$  such that  $\mathbb{P}(\mathcal{R}(\hat{c}) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) \leq \epsilon) \geq 1 - \delta$ .

•  $\mathcal{A}$  is called an efficient PAC algorithm if it runs in polynomial of  $\frac{1}{\epsilon}$  and  $\frac{1}{\delta}$ .

•  $\mathcal{C}$  is (efficiently) PAC-learnable from  $\mathcal{H}$  if there is an algorithm  $\mathcal{A}$  that (efficiently) learns  $\mathcal{C}$  from  $\mathcal{H}$ .

• Finite  $\mathcal{C}$ ,  $\mathbb{P}(\mathcal{R}(\hat{c}_n) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) > \epsilon) \leq 2|\mathcal{C}| \exp(-\frac{n\epsilon^2}{2})$  is PAC-learnable.

•  $\mathcal{C}$  with  $\dim_{VC} = d < \infty$  is PAC-learnable,  $\mathbb{P}(\mathcal{R}(\hat{c}_n) - \inf_{c \in \mathcal{C}} \mathcal{R}(c) > \epsilon) \leq 9n^d \exp(-\frac{n\epsilon^2}{32})$

## A Appendix

(1)  $\partial_x(AB) = A \partial_x B + (\partial_x A) B$ , (2)  $\partial_x A^{-1} = -A^{-1} (\partial_x A) A^{-1}$ , (3)  $\partial_x \ln \det A = \text{Tr}(A^{-1} \partial_x A)$ ,

Define  $(\partial_A f)_{ij} := \partial_{a_{ij}} f$ , then (4)  $\partial_A \text{Tr}(BA) = \partial_A \text{Tr}(AB) = B$ , (5)  $\partial_A \ln \det A = A^{-1}$ , (6)  $\partial_A \text{Tr}(ABA^\top) = (B + B^\top) A^\top$ ,

$\mathcal{N}(\mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-(x-\mu)^\top \Sigma^{-1} (x-\mu)/2}$ , Conditional  $\mathbb{E}[y_2 | y_1] = \mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (y_1 - \mu_1)$ ,  $\text{Cov}[y_2 | y_1] = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$ . Marginal  $\mathbb{E}(y_2) = \mu_2$ ,  $\text{Cov}[y_2] = \Sigma_{22}$

$$(A + UC^{-1}V)^{-1} = A^{-1} - A^{-1}U(C + VA^{-1}U)^{-1}VA^{-1}.$$