1 Basics

$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$, $\mathcal{N}(x|\mu,\sigma)$

$f(x) = \frac{1}{\sqrt{(2\pi)^d \det\Sigma}} e^{-\frac12 (x-\mu)^T\Sigma^{-1}(x-\mu)}$, $\mathcal{N}(x|\mu,\Sigma)$

$X\sim\mathcal{N}(\mu,\Sigma), Y=A+BX \Rightarrow Y\sim\mathcal{N}(A+B\mu, B\Sigma B^T)$

$log(\mathcal{N}(x|\mu,\Sigma)) = \frac12 log|\Sigma^{-1}| - \frac12(x-\mu)^T\Sigma^{-1}(x-\mu) + const \Rightarrow \frac{\partial log \mathcal{N}(x|\mu,\Sigma)}{\mu} = \Sigma^{-1}(x-\mu)$, $\frac{\partial log\mathcal{N}(x|\mu,\Sigma)}{\Sigma^{-1}} = \frac12\Sigma - \frac12(x-\mu)(x-\mu)^T$

f(x) on a: $f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + ...$

$p(\begin{bmatrix}a_1\\a_2\end{bmatrix}) = \mathcal{N}(\begin{bmatrix}u_1\\u_2\end{bmatrix}, \begin{bmatrix}\Sigma_{11}\Sigma_{12}\\ \Sigma_{21}\Sigma_{22}\end{bmatrix})$, $p(a_2|a_1) = \mathcal{N}(u_2 + \Sigma_{21}\Sigma_{11}^{-1}(a_1-u_1), \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$

- $Var[X] = \int_x (x-\mu)^2 p(x)dx$
- $Var[aX] = a^2 Var[X]$
- $Var[X] = E[(X-E[X])^2] = E[X^2]-E[X]^2$
- $Var[X+Y] = Var[X]+Var[Y]+2Cov[X,Y]$
- $Cov[X,Y] = E[(X-E[X])(Y-E[Y])]$
- $Cov[aX,bY]=abCov[X,Y]$ • $\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top\mathbf{x}) = \frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{b}) = \mathbf{b}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x}$
- $\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = (\mathbf{A}^\top+\mathbf{A})\mathbf{x} \overset{\text{A sym.}}{=} 2\mathbf{A}\mathbf{x}$
- $\frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{A}\mathbf{x}) = \mathbf{A}^\top\mathbf{b}$ • $\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{X}\mathbf{b}) = \mathbf{c}\mathbf{b}^\top$
- $\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{X}^\top\mathbf{b}) = \mathbf{b}\mathbf{c}^\top$ • $\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}-\mathbf{b}\|_2) = \frac{\mathbf{x}-\mathbf{b}}{\|\mathbf{x}-\mathbf{b}\|_2}$
- $x^TAx = Tr(x^TAx) = Tr(xx^TA) = Tr(Axx^T)$
- $\frac{\partial}{\partial A}Tr(AB) = B^T$ • $\frac{\partial}{\partial A}log|A| = A^{-T}$
- sigmoid$(x) = \sigma(x) = \frac{1}{1+\exp(-x)}$
- ∇sigmoid(x) = sigmoid$(x)(1-$ sigmoid$(x))$
- $CE(y,\hat y) = -(y log\hat y + (1-y)\log(1-\hat y))$

2 Anormaly Detection

Dimensionality Reduction: Simpler case: d = 1 ($\pi: R^D\to R$); Assume $\pi(X) = u_1X$ with $\|u_1\|^2 = 1$: Mean of proj. data: $u_1^T\overline{X}$ ($\overline{X} = \frac1n\sum_{x\in X}x$); Variance of proj. data: $\frac1n\sum_{i\le n}(u_1^T\overline{X} - u_1^Tx_i)^2 = u_1^TCov(X)u_1 := u_1^TSu_1$

Objective: $max_{u_1\in R^D} u_1^TSu_1$ s.t. $\|u_1\|^2=1$

Lagrangian: $\mathcal{L}(u_1) = u_1^TSu_1 + \lambda(1-u_1^Tu_1)$; $\frac{\partial\mathcal{L}}{\partial u_1} = 0 \Rightarrow Su_1 = \lambda u_1$; $u_1^*Su_1^* = \lambda$

GMM: $max_{\pi_k,\mu_k,\Sigma_k} log(p(x)) = log(\sum_k \pi_k\mathcal{N}(x|\mu_k,\Sigma_k))$ s.t. $\sum_{k=1}^K\pi_k = 1$, Σ_k is p.d.; $log p_\theta(X) = \mathbb{E}_{z\sim q}[log p_\theta(X)] = \mathbb{E}_z[log p_\theta(X,z)] - \mathbb{E}_z[log q(z)] + \mathbb{E}_z[log(\frac{q(z)}{p_\theta(z|X)})] := M(q,\theta) + E(q,\theta)$ (intractable).

EM: Properties: 1) $E(q,\theta)\ge 0$, $M(q,\theta)\le log p_\theta(X)$ 2) $E(q^*,\theta) = 0$ for $q^* = min_q E(q,\theta) = p_\theta(z|X)$, $M(q^*,\theta) = logp_\theta(X)$ 3) $logp_\theta(X) = M(q,\theta)+E(q,\theta) = M(q^*,\theta) + 0 = \mathbb{E}_{z\sim q^*}[logp_\theta(X,z)] - \mathbb{E}_{z\sim q^*}[logq(z)] \le max_\theta M(q^*,\theta)$; E-step: $q_t^* = min_q E(q,\theta^t)$; M-step: $\theta^{(t+1)} = max_\theta M(q_t^*,\theta)$

3 Density Estimation

Fisher Info & Cramér–Rao Bound: $E_X[(\theta-\hat\theta)^2] \ge \frac{(\frac{\partial}{\partial\theta}bias(\hat\theta)+1)^2}{I_n(\theta)} + bias^2(\hat\theta)$, where $I_n(\theta) = nI_1(\theta) = \mathbb{E}[(\frac{\partial}{\partial\theta}logp(x|\theta))^2] = \mathbb{E}[\wedge^2]$, $\wedge := \frac{\frac{\partial}{\partial\theta}p(x|\theta)}{p(x|\theta)}$; Properties: 1) $\mathbb{E}_X[\wedge] = \int p(X|\theta)\frac{\frac{\partial}{\partial\theta}p(x|\theta)}{p(x|\theta)}dX = \frac{\partial}{\partial\theta}\int p(X|\theta)dX = 0$ 2) $\mathbb{E}_X[\wedge\hat\theta] = \frac{\partial}{\partial\theta}\int p(X|\theta)\hat\theta(X)dX = \frac{\partial}{\partial\theta}\mathbb{E}_X[\hat\theta] = \frac{\partial}{\partial\theta}bias(\hat\theta)+1$

Proof: $Cov(\wedge,\hat\theta) = \mathbb{E}_X[(\wedge - \mathbb{E}_X[\wedge])(\hat\theta - \mathbb{E}_X[\hat\theta])] = \mathbb{E}[\wedge\hat\theta] - \mathbb{E}[\wedge]\mathbb{E}[\hat\theta] = \frac{\partial}{\partial\theta}bias(\hat\theta)+1$

$Cov(\wedge,\hat\theta)^2 \le \mathbb{E}_X[(\wedge - \mathbb{E}_X[\wedge])^2]\mathbb{E}_X[(\hat\theta - \mathbb{E}_X[\hat\theta])^2] = \mathbb{E}_X[\wedge^2]\mathbb{E}_X[(\hat\theta-\theta-\mathbb{E}[\hat\theta]+\theta)^2] = \mathbb{E}[\wedge^2](\mathbb{E}[(\hat\theta-\theta)^2] - bias^2(\hat\theta))$

Approaches: Frequentism (MLE): Desiderata, asymptotically unbiased but large variance (out-performed by biased estimators, e.g. shrinked estimators and Stein's).

Frequentism fulfills the desiderata: 1) Asymptotic Efficiency: $lim_{n\to\infty}\mathbb{E}[(\hat\theta-\theta)^2] = \frac{1}{I_n(\theta)}$ 2) Consistency $lim_{n\to\infty}p(|\hat\theta_n - \theta|>\epsilon) = 0, \forall\epsilon>0$ 3) Asymptotic Normality $\hat\theta_n \to \mathcal{N}(\theta,\sigma^2), \sigma>0$

Bayesianism: Prior induces a regularization effect that raises the bias but decreases the variance; To avoid intractability issues, use conjugate priors.

Statistical Learning: tractable with low bias and variance, but hard to select model.

Logistic Regression, Frequentism: $\mathbb{E}[y|X=x,\hat\theta] = p(y=1|X=x,\hat\theta) = \frac{p(X=x|y=1,\hat\theta)p(y=1|\hat\theta)}{p(X=x|y=1,\hat\theta)p(y=1|\hat\theta)+p(X=x|y=0,\hat\theta)p(y=0|\hat\theta)} = \frac{1}{1+\frac{p(X=x|y=0,\hat\theta)}{p(X=x|y=1,\hat\theta)}} = \frac{1}{1+exp(-w^Tx+w_0)} = \sigma(w^Tx+w_0)$

LR, Bayesianism: Prior: $p(w) = \mathcal{N}(w|m_0,S_0) = \mathcal{N}(w|0,\alpha I)$; Likelihood: $p(X,y|w) = \prod_i\sigma(x_i^Tw)^{y_i}(1-\sigma(x_i^Tw))^{1-y_i}$; intractable

Approximate by Laplace's Method: $p(w|x,y) = \frac{p(w,x,y)}{p(x,y)} \propto exp(-(-logp(w,x,y))) := exp(-(R(w)))$; $R(w)\approx R(w^*) + \cancel{(w-w^*)^T\nabla R(w^*)} + \frac12(w-w^*)^TH_R(w-w^*)$, with $w^* = min_wR(w)$ $\Rightarrow p(w|x,y)\approx\mathcal{N}(w|w^*, H_R^{-1}(w^*))$

LR, Statistical Learning: Model: $\mathcal{H} = \{f|f: R^d\to[0,1], f(x) = \sigma(w^Tx)\}$

Loss function: $\mathcal{L}(y,f(x)) = -logp_{f(x)}(y)$; Expected loss: $\mathbb{E}_{X,y\sim p^*}[\mathcal{L}(y,f(X))] = \mathbb{E}_X\mathbb{E}_{y|X}[-logp_{f(x)}(y)]$; Empirical loss: $\frac1n\sum_{i\le n}(-y_ilog\sigma(w^Tx_i) - (1-y_i)log(1-\sigma(w^Tx_i)))$ (same as frequentist approach)

BIC: for $S\subseteq\{1,...,d\}$, $\mathcal{H}_S = \{f: R^{|S|}\to[0,1]\}$; When $p(w) = \mathcal{N}(w|m_0,\alpha_0I)$ (for large α_0), $logp(x,y)\approx logp(w^*) + log(x,y|w^*) - \frac{|S|}{2}log(2\pi) - \frac12 log|H_R| \approx const - \frac12(|S|logn - 2logp(x,y|w^*))$; Lower BIC, better model.

4 Regression

Linear Regression: $RSS(\beta) = \sum_{i=1}^n(y_i - x_i^T\beta)^2 = (y-X\beta)^T(y-X\beta)\Rightarrow\hat\beta = (X^TX)^{-1}X^Ty$

Prove $\hat\beta$ is unbiased: $\hat\theta := a^T\hat\beta$, $\mathbb{E}_\epsilon[\hat\theta] = \mathbb{E}_\epsilon[a^T(X^TX)^{-1}X^Ty] = a^T(X^TX)^{-1}X^T\mathbb{E}_\epsilon[y] = a^T(X^TX)^{-1}X^T(X\beta+\mathbb{E}_\epsilon[\epsilon]) = a^T\beta$

Alternative unbiased estimator: $\tilde\theta = c^Ty = a^T\hat\beta + a^TDy = a^T\beta + a^TDX\beta = a^T\beta$; $a^TDX = 0$

Gauss Markov Theorem: $\forall\tilde\theta = c^Ty$ unbiased for $a^T\hat\beta, \mathbb{V}(a^T\hat\beta)\le\mathbb{V}(c^Ty)$; Proof: $\mathbb{V}(c^Ty) = \mathbb{E}[(c^Ty)^2] - \mathbb{E}[c^Ty]^2 = c^T(\mathbb{E}[yy^T] - \mathbb{E}[y]\mathbb{E}[y]^T)c = \sigma^2c^Tc = \sigma^2(a^T(X^TX)^{-1}a + a^TDD^Ta) = \mathbb{V}(a^T\hat\beta) + \sigma^2a^TDD^Ta$

Bias-variance Tradeoff: $\mathbb{E}_D\mathbb{E}_{Y|X=x}(\hat f(x) - Y)^2 = \mathbb{E}_D(\hat f(x) - \mathbb{E}(Y|X=x))^2 + \mathbb{E}(Y-\mathbb{E}(Y|X=x))^2 = \mathbb{E}_D(\hat f(x) - \mathbb{E}_D\hat f(x))^2 + (\mathbb{E}_D\hat f(x) - \mathbb{E}(Y|X=x))^2 + \mathbb{E}(Y-\mathbb{E}(Y|X=x))^2$ = var + bias^2 + noise

Regularization: Can be viewed as MAP estimation with a prior. Ridge: $\beta\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda I})$; Lasso: $p(\beta_i) = \frac{\lambda}{4\sigma^2}exp(-|\beta|\frac{\lambda}{2\sigma^2})$ (Laplace, no closed-form solution since l_1 norm is not differentiable, more sparse estimations since the gradient of regularization does not shrink as Ridge)

Bayesian LR: Assume $\epsilon\sim\mathcal{N}(0,\sigma^2I)$, $\beta\sim\mathcal{N}(0,\wedge^{-1})$, $p(\beta|Y,X,\sigma^2,\wedge) = \mathcal{N}((X^TX+\sigma^2\wedge)^{-1}X^TY, \sigma^2(X^TX+\sigma^2\wedge)^{-1})$

Bayesian LR is a special case of Gaussian Processes with linear kernel $k(x,x') = x^T\wedge^{-1}x'$

5 Gaussian Processes

Prediction with GP: $p(y_{n+1}|x_{n+1},X,y) = \mathcal{N}(k(x_{n+1},X)^T(k(X,X)+\sigma^2I)y, k(x_{n+1}x_{n+1}) + \sigma^2 - k(x_{n+1},X)^T(k(X,X)+\sigma^2I)^{-1}k(x_{n+1},X))$

Kernels: Properties: 1) $k(x,x') = k(x',x)$ 2) $x^TKx\ge 0\forall x$ 3) $k(x,x') = \phi(x)\phi(x')$; Composition: addition, multiplication, scaling, $k(x,x') = f(k_1(x,x')) = f(x)k_1(x,x')f(x')$ for positive polynomial or exponential f; Constructions: 1) $k(x,x') = k_1(x,x')+k_2(x,x')$, Proof: ∃ symmetric gram matrices K_1,K_2 s.t. $x^TK_1x, x^TK_2x\ge0\Rightarrow x^TKx = x^T(K_1+K_2)x\ge 0$ 2) $k(x,x') = k_1(x,x')k_2(x,x')$, Proof: $k_1(x,x')k_2(x,x') = \Sigma_{i,j}(f_i(x)g_j(x))(f_i(x')g_j(x')) = \Sigma_{i,j}h_{i,j}(x)h_{i,j}(x') = \phi(x)\phi(x')$ 3) $k(x,x') = exp(k_1(x,x'))$, Proof: $\sum_{i=1}^m\frac{k_1(x,x')^r}{r!}\to k(x,x')$ as m → ∞ 4) RBF: $k(x,x') = exp(-\frac{1}{2\gamma^2}\|x-x'\|_2^2) = exp(-\frac{1}{2\gamma^2}\|x\|_2^2)exp(\frac{1}{\gamma^2}x^Tx')exp(-\frac{1}{2\gamma^2}\|x'\|_2^2)$ (larger bandwidth $\gamma\to$ smoother curves)

6 Ensemble Methods

Bagging: $\mathbb{E}[(y-b^{(M)}(x))^2] = bias^2(b^{(M)}(x)) + var(b^{(M)}(x)) = bias^2(b(x)) + \frac1M var(b(x)) \le \mathbb{E}[(y-b(x))^2]$; Random Forests chooses m random features at each splitting step (i.d. base models). Randomized feature selection induces implicit regularization; no overfitting

AdaBoost: $b^{(0)} = 0, w_i^{(0)} = \frac1n$; 1) $b^{(t)} = min_\beta\Sigma_iw_i^{(t)}\mathbb{I}\{b(x_i)\ne y_i\}$ 2) Evaluate err_t 3) $\tilde\alpha_t = \frac12log(\frac{1-err_t}{err_t})$, $b^{(t)} = b^{(t-1)} + \tilde\alpha_tb^{(t)}$ 4) $w_i^{(t+1)} = w_i^{(t)}exp(-\tilde\alpha_ty_ib^{(t)})$ 5) Renormalize $w^{(t+1)}$; Output $\sum(\tilde\alpha_i(b^i(x))$

Forward stagewise additive modeling: Proof of 3): $\mathbb{E}[f(x)] := \mathbb{E}[exp(-yf(x))] = P(Y=1|X=x)exp(-f(x)) + P(Y=-1|X=x)exp(f(x))$ $\frac{\partial\mathbb{E}[f(x)]}{\partial f(x)} = 0 \Rightarrow f^*(x) = \frac12\frac{P(Y=1|X=x)}{P(Y=-1|X=x)}$; Proof of 1): $min_{\alpha>0,b\in\mathcal{H}}\sum_i\mathcal{L}(y_i,\alpha b(x_i)+f_{t-1}(x_i)) = min_{\alpha,b}\sum_iw_i^{(t)}exp(-\alpha y_ib(x_i)) = \min_{\alpha,b}\sum_{i,y_i\ne b(x_i)}w_i^{(t)}e^\alpha + (\sum_iw_i^{(t)}e^{-\alpha} - \sum_{i,y_i\ne b(x_i)}w_i^{(t)}e^{-\alpha})$; $w_i^{(t)} = exp(-y_if_{(t-1)}(x_i))$

Gradient Boosting: $\hat f_0(x) = min_h\Sigma_{i=1}^n(y_i - h(x_i))^2$; 1) $g_t(x_i) = [\frac{\partial\mathcal{L}(y_i,f(x_i))}{\partial f(x_i)}]_{f=\hat f_{t-1}(x_i)}$ 2)$h_t = min_h\Sigma_i(-g_t(x_i) - h(x_i))$ 3)$\beta_t = min_\beta\Sigma_i\mathcal{L}(y_i,\hat f_{t-1}(x_i)+\beta h_t(x_i))$ 4)$\hat f_t(x) = \hat f_{t-1} + \beta_th_t(x)$; Output $\hat f_t$;

7 Convex Optimization & SVMs

Duality: Primal: $min_\omega f(\omega)$ s.t. $g_i(\omega) = 0$ and $h_j(\omega)\le 0$ Dual: $max_{\lambda,\alpha}\theta(\lambda,\alpha)$ s.t. $\alpha_j\ge0$

Weak duality: $\theta(\lambda,\alpha) = inf_{\omega\in\mathbb{R}^d}\mathcal{L}(\omega,\lambda,\alpha\ge0) \le \mathcal{L}(\omega^*,\lambda,\alpha) = f(\omega^*) + \sum_i\lambda_ig_i(\omega^*)$ (= 0) $+\sum_j\alpha_jh_j(\omega^*)(\le0)\le f(\omega^*)$

Slater's condition (check if strong duality holds): $\exists\omega$ s.t. $g_i(\omega) = 0, h_i(\omega) < 0\ \forall i,j$

Strong Duality (if Slater's holds, convex f, nonconvex g, linear h): 1) $\omega^* = min_\omega \mathcal{L}(\omega, \lambda^*, \alpha^*)$ 2) Complementary slackness: $\alpha_j h_j(\omega^*) = 0, \forall j$

Linearly separable SVM: Primal: $max_{w,w_0} 2m(w,w_0) = \frac{|w^Tx^+ - w^Tx^-|}{\|w\|} = max_{w,w_0}\frac{2}{\|w\|} = min_{w,w_0}\frac{1}{2}\|w\|^2$ for random x^+, x^-; $y_i(w^Tx_i+w_0)\geq 1, \forall i$
Slater's: take $(\gamma w, \gamma w_0)$, $\gamma y_i(w^Tx_i+w_0) > 1$
Dual: $\theta(\alpha) = min_{w,w_0}\mathcal{L}(w,w_0,\alpha)$ s.t. $\alpha_i \geq 0, \forall i$ $= min_{w,w_0}\frac12\|w\|^2 + \sum_i \alpha_i(1-y_i(w^Tx_i+w_0)) \Leftrightarrow max_\alpha -\frac12\sum_{ij}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_i\alpha_i$ $(\sum_i \alpha_i y_i = 0$; $w^* = \sum_i \alpha_i y_i x_i$; $w_0^* = -\frac12(w^{*T}x^+ - w^{*T}x^-))$
Compl. slack.: $\alpha_i^*(1-y_i(w^{*T}+w_0^*)) = 0 \Rightarrow \alpha_i^* = 0 \Rightarrow w^*$ is a sparse comb. of support vectors

Linearly inseparable SVM: Primal: $min_{w,w_0,\xi}\frac12\|w\|^2 + C\sum_i\xi_i$; $y_i(w^Tx_i+w_0)\geq 1-\xi_i$; $\xi_i\geq0$; larger C means narrower margin, fewer neglected samples, and fewer support vectors.
Dual: $L(w,w_0,\xi,\alpha,\beta) = \frac12 w^Tw + C\sum_{i=1}^n\xi_i - \sum_{i=1}^n\beta_i\xi_i - \sum_{i=1}^n\alpha_i(y_i(w^T\phi(x_i)+w_0)-1+\xi_i)$; $0\leq\alpha_i\leq C$; $\xi_i^* = max(0, 1-y_i(w^{*T}x_i+w_0^*))$

Kernelization: Dual: $max_\alpha -\frac12\sum_{ij}\alpha_i\alpha_jy_iy_j\phi(x_i)^T\phi(x_j)+\sum_i\alpha_i$; $w^{*T}\phi(x) = \sum_i\alpha_i^*y_i\cancel{\phi(x_i)^T\phi(x)}k(x_i,x)$

Extensions: SVM Regression: ϵ-sensitive loss: $max(0,|y-f(x)|-\epsilon)$; Primal: $min_{w,\xi,\hat\xi}\|w\|^2 + C\sum_i(\xi_i+\hat\xi_i)$ s.t. $(w^Tx_i+w_0)-y_i\leq\epsilon+\xi_i$, $y_i-(w^Tx_i+w_0)\leq\epsilon+\hat\xi_i$, $\xi_i,\hat\xi_i\geq 0$; Dual: $max_{\alpha,\hat\alpha}\sum_i(\hat\alpha-\alpha)y_i - \epsilon\sum_i(\hat\alpha+\alpha) - \frac12\sum_{i,j}(\hat\alpha_i-\alpha_i)(\hat\alpha_j-\alpha_j)x_ix_j$ s.t. $0\leq\alpha_i,\hat\alpha_i\leq C$, $\sum_{i,j}(\hat\alpha_i-\alpha_i)=0, \forall i$; Multi-class SVM: Constraint: $\forall y\in\{1,...,M\}, \forall x_i\in X$, $(w_{y_i}^Tx_i+w_{y_i,0}) - max_{y\neq y_i}(w_y^Tx_i+w_{y,0})\geq 1-\xi_i$; Structural SVM: Constraint: $w^T\Phi(y_i,x_i) - max_{y\neq y_i}[\Delta(y,y_i)+w^T\Phi(y,x_i)]\geq -\xi_i, \forall x_i\in X$

8 Deep Learning & Generative Models

Robbins-Monro Method: $X_{n+1} = X_n - \alpha_n(f(x_n)+\gamma_n)$; Conditions: 1) $lim_{n\to\infty}\alpha_n = 0$ (convergence) 2) $\sum_{n=1}^\infty\alpha_n=\infty$ (slow enough to find root) 3) $\sum_{n=1}^\infty\alpha_n^2<\infty$ (bounded variance); Proof: $x_{n+1}-x_0 = x_n - x_0 - \alpha_n(f(x_n)+\gamma_n) \Leftrightarrow \mathbb{E}[(x_{n+1}-x_0)] = \mathbb{E}[(x_n-x_0)] - 2\alpha_n\mathbb{E}[(x_{n+1}-x_0)(f(x_n)+\gamma_n)] + \alpha_n^2\mathbb{E}[f^2(x_n)+2f(x_n)\gamma_n+\gamma_n^2] = \mathbb{E}[(x_n-x_0)]+\alpha_n^2\mathbb{E}[(x_n-x_0)f(x_n)]-2\alpha_n\mathbb{E}[\gamma_n^2]$; Iterate n-1 times to reduce x_n s.t. $\mathbb{E}[(x_{n+1}-x_0)] - \mathbb{E}[(x_1-x_0)]\leq(b+\sigma^2)\sum_{i=1}^{n-1}\alpha_i^2 - 2\sum_{i=1}^{n-1}\alpha_i\mathbb{E}[(x_i-x_0)f(x_i)]$; LHS bounded from below & RHS $\to -\infty$ iff $(x_i-x_0)f(x_i)\geq 0 \Rightarrow lim_{n\to\infty}P(x_n = x_0) = 1$.

Optimality for step size: $f(x_n+\Delta x) = f(x_n) + \nabla f(x_n)^T\Delta x + \frac12\Delta x^T H\Delta x$; Since $\Delta x = x_{n+1}-x_n = -\alpha_n\nabla f(x_n)$, $f(x_{n+1}) = f(x_n) - \alpha_n\nabla f(x_n)^T\nabla f(x_n) + \frac12\alpha_n^2\nabla f(x_n)^TH\nabla f(x_n)$; Assume $\frac{\partial}{\partial\alpha_n}f(x_{n+1}) = \frac{\partial}{\partial\alpha_n}f(x_0) = 0 \Leftrightarrow \alpha_n = \frac{\nabla f^T\nabla f}{\nabla f^TH\nabla f} = H^{-1}$

SGD: Nesterov Momentum: $y_{n+1} = x_n+\beta(x_n-x_{n-1})$; $x_{n+1} = y_{n+1}-\alpha_n\nabla f(y_{n+1})$ for $\beta>0$; SGD with Momentum: $x_{n+1} = y_{n+1}-\alpha_n\nabla f_{I(n)}(y_{n+1})$ with $I(n)\sim$ Unif$\{1,...,n\}$; Sign SGD: $x_{n+1} = x_n - \alpha_n sign(\nabla f_{I(n)}(x_n))$; Mini-batch: $x_{n+1} = x_n - \alpha_n\frac1B\sum_{i\in B}^n\nabla f_i(x_n)$; Unbiased grad: $\mathbb{E}_{I(n)}[\nabla f_{I(n)}] = \frac1n\sum_{i=1}^n\nabla f_i(x) = \nabla f(x)$

VAEs: Problem: $max_\theta p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz$ is intractable; Solution: define encoder $q_\theta(z|x)$ that approximates $p_\theta(z|x)$; $logp_\theta(x) = \mathbb{E}_{z\sim q_\theta(z|x_i)}[logp_\theta(x_i)] = \mathbb{E}_z[log\frac{p_\theta(x_i|z)p_\theta(z)}{p_\theta(z|x_i)}\frac{q_\theta(z|x_i)}{q_\theta(z|x_i)}] = \mathbb{E}_z[logp_\theta(x_i|z)] - \mathbb{E}_z[log\frac{q_\theta(z|x_i)}{p_\theta(z)}] + \mathbb{E}_z[log\frac{q_\theta(z|x_i)}{p_\theta(z|x_i)}] =$
$\mathbb{E}_z[log\frac{q_\theta(z|x_i)}{p_\theta(z)}] - KL(q_\theta(z|x_i)\|p_\theta(z))$
$+KL(q_\theta(z|x_i)\|p_\theta(z|x_i))\geq\mathcal{L}(x_i;\theta,\phi)$ (ELBO)

HVAEs: Hierarchical latent vectors, top-down shared model with learnable mean and variance to keep long-range data correlations and avoid posterior collapse

GANs: Objective: $min_G max_D$ $\{\mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p(z)}[log(1-D(G(z))]\}$
This loss is essentially 2 KL divergences. At early stages, the 2 distributions don't overlap substantially, which leads to vanishing gradient. Solution: Wasserstein Distance $WG_r(p_1,p_2) = (\mathbb{E}_{x\sim p_1,y\sim p_2}[\|x-y\|^r])^{\frac1r}$

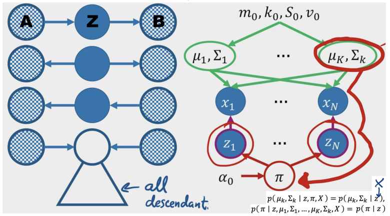
Extracting representations invariant from domains: Conditional GANs: $D\subseteq W\times X\times y$ 1) $E: X\to Z$ 2) $F: Z\to[0,1]^y$ 3) $D: Z\times y\to[0,1]^W$ Objective: $min_{E,F}\ max_D\ \mathbb{E}_X[CE(p(y|x),\hat p_{E(X)}(\cdot))] - \lambda\mathbb{E}_{X,y}[CE(p_{w|x,y},\hat p_{E(X),y}(\cdot))]$
Maximum-mean discrepency: Goal: view representations from 2 domains as 2 samples from the same distribution. $MMD(p,q) = sup_{f\in\mathcal{F}}(\mathbb{E}_{X\sim p}[f(X)] - E_{y\sim q}[f(y)])^2 \approx sup_{f\in\mathcal{H}_0}(\sum_i w_i\mathbb{E}[x_i] - \sum_i w_i\mathbb{E}[y_i]) = sup_{f\in\mathcal{H}_0}(w^T(\mathbb{E}[x^i]) - w^T(\mathbb{E}[y^i])) = sup_{f\in\mathcal{H}_0}\langle f,\mu_p-\mu_q\rangle = \|\mu_p-\mu_q\|^2 = \mathbb{E}[\sum_i\phi_i(x_1)\phi_i(x_2) - 2\sum_i\phi_i(x)\phi_i(y) + \sum_i\phi_i(y_1)\phi_i(y_2)]$; Objective: $min_{E,F}\mathcal{L}_C(E,F)+\lambda\hat{\mathcal{L}}_{MMD}(E)$

Diffusion Models: $Z_i = \beta_iz_{i-1}+\beta_i\epsilon$, $\epsilon\sim\mathcal{N}(0,I)$, $q(z_i|z_{i-1}) = \mathcal{N}(z_i|\beta_iz_{i-1},\beta_iI)$; $q(z_t|x) = q(z_t|z_{t-1})...q(z_1|x) = \mathcal{N}(z_t|\beta_tz_{t-1},\beta I)...\mathcal{N}(z_t|\beta_1x,\beta_1I) = \mathcal{N}(z_t|\sqrt{\tilde\alpha_t}x,(1-\tilde\alpha_t)I)$, where $\tilde\alpha_t = \prod_s(1-\beta_s)$; Forward posterior: $q(z_{t-1}|z_t,x) = \mathcal{N}(z_{t-1}|\tilde\mu_t(z_t,x),\tilde\beta_tI)$, where $\tilde\mu_t = \frac{\sqrt{\tilde\alpha_{t-1}}\beta_t}{1-\tilde\alpha_t}x + \frac{\sqrt{1-\beta_t}(1-\tilde\alpha_{t-1})}{1-\tilde\alpha_t}z_t$, $\tilde\beta_t = \frac{1-\tilde\alpha_{t-1}}{1-\tilde\alpha_t}\beta_t$; New ELBO: $\mathbb{E}[logp(x|z_1)] - KL(q(z_n|x)\|p(z_n)) - \sum_i\mathbb{E}[KL(q(z_{i-1}|z_i,x)\|p(z_{i-1}|z_i))]$

9 Non-parametric Bayesian Inference

BI for multivariate Gaussian: $p(x^*|X) = \int p(x^*|\theta)p(\theta|X)d\theta = \mathbb{E}_{\theta\sim p(\cdot|X)}[p(x^*|\theta)] \approx \frac1M\sum_tp(x^*|\theta^{(t)})$ where $\theta^{(t)}\sim p(\cdot|X)$; $\mu\sim\mathcal{N}(m_0,V_0),\Sigma\sim IW(S_0,v_0)\Rightarrow\mu|\Sigma,X\sim\mathcal{N}(m_p,V_p),\Sigma|\mu,X\sim IW(S_p,v_p)$ Gibbs sampling: For semi-conjugate priors, iteratively resample acc. to tractable cond. dist. n times. The update does not need to be in exact order for 1-dim and first M samples are discarded.

BI for GMM: Dirichlet distribution (DP) on π: $Dir(\pi|\alpha) = \frac{\Gamma(\sum_i\alpha_i)}{\prod_i\Gamma(\alpha_i)}\prod_i\pi_i^{\alpha_i-1}$, $\sum_i\pi_i=1$



If every path from variable A to B is blocked by d-separation Z, then A and B are independent conditioned on Z.
Collapsed Gibbs sampling: first sample z: $p(z_i=k|z_{-i}=\zeta,X)\propto p(z_i=k|z_{-i}=\zeta)p(X|z_i=k,z_{-i}=\zeta)\propto p(z_i=k|z_{-i}=\zeta)p(x_i|X_{-i},z_i=k,z_{-i}=\zeta)p(X_{-i}|z_i=k,z_{-i}=\zeta)\propto p(z_i=k|z_{-i}=\zeta)p(x_i|\{x_j:j\leq N_{i\neq j},z_j=k\})const$;
Rao-Blackwellization: $Var_Z[\mathbb{E}_\theta[f(\theta,Z)|Z]] = \mathbb{E}_Z[(\mathbb{E}_{\theta,Z}[f(\theta,Z)] - \mathbb{E}_{\theta'}[f(\theta',Z)])^2] = \mathbb{E}_Z[(\mathbb{E}_{\theta'}[\mathbb{E}_{\theta,Z}[f(\theta,Z)] - f(\theta',Z)])^2] \leq \mathbb{E}_Z[\mathbb{E}_{\theta'}[(\mathbb{E}_{\theta,Z}[f(\theta,Z)] - f(\theta',Z))^2]] = \mathbb{E}_{Z,\theta'}[(\mathbb{E}_{\theta,Z}[f(\theta,Z)] - f(\theta',Z))^2] = Var_{\theta',Z}[f(\theta',Z)]$

BI for non-parametric GMMs: Sampling prior: 1) Draw π from $GEM(\alpha)$ with Stick-breaking Process: $\pi_1=\beta_1\sim Beta(1,\alpha)$, $\pi_{i,i\geq2} = \prod_{j<i}(1-\beta_j)\beta_i$; 2) Chinese Restaurant Process (metaphor of DP, draw z directly): $p(z_n=k) = \begin{cases} n_k/(\alpha+n-1), \text{for existing k}\\ \alpha/(\alpha+n-1), \text{for leftmost empty k}\end{cases}$
$z_1,...,z_n$ are not independent but exchangable. Proof: $p(z_1=k_1,...,z_n=k_n) = \prod_ip(z_i=k_i|z_1=k_1,...,z_n=k_n) = \prod_i\frac{f(\alpha,k_i)}{\alpha+i-1} = \prod_i\frac{f(\alpha,k_{\pi^{-1}(i)})}{\alpha+i-1} = p(z_{\pi(1)}=k_1,...,z_{\pi(n)}=k_n)$; Asymptotics of the expected # of distinct samples drawn / expected # of occupied tables in CRP: $S(n) = \sum_k\frac{\alpha}{\alpha+k-1}\geq I(n) = \int_1^{n+1}\frac{\alpha}{\alpha+x-1}dx = \alpha(ln(\frac{\alpha+n}{\alpha}))$
DeFinetti's Theorem: any exchangeable distribution admits a mixture model, $p(X_1=x_1,...,X_n=x_n) = \int\prod_ip(x_i|\theta)p(\theta)d\theta$

10 PAC Learning

Algorithm $\mathcal{A}$ can learn $c\in\mathcal{C}$ if $\exists poly(\cdot,\cdot,\cdot)$, s.t. (1) $\forall$ dist. D on X and (2) $\forall\epsilon\in[0,\frac12],\delta\in[0,\frac12]$, $\mathcal{A}$ outputs $\hat c\in H$ given a sample of size at least $poly(\frac1\epsilon,\frac1\delta,size(c))$ s.t. $p_{Z\sim D^n}(\mathcal{R}(\hat c)\leq\epsilon+inf_{c\in\mathcal{C}}\mathcal{R}(c))\geq1-\delta$; $\mathcal{A}$ is an efficient PAC algorithm if it runs in polynomial of $\frac1\epsilon$ and $\frac1\delta$. $\mathcal{C}$ is (efficiently) PAC-learnable from $\mathcal{H}$ if there is an algorithm $\mathcal{A}$ that learns $\mathcal{C}$ from $\mathcal{H}$.

Rectangle Problem: $n\geq\frac4\epsilon ln\frac4\delta$, suffices to prove $p(\mathcal{R}(\hat R)\leq\epsilon)\geq p(\hat RIG)\geq 1-4exp(-\frac{n\epsilon}{4})$; Proof: $p(\neg\hat RIG)\leq\sum_l\prod_ip(x_i\notin T_l^\epsilon) = 4(1-\frac\epsilon4)^n\leq4exp(-\frac{n\epsilon}4)$; Generalization: for $n\geq\frac1\epsilon(log|\mathcal{H}|+log\frac1\delta)$, $\hat R(\hat h)=0\Rightarrow$ 1) prove $|\mathcal{H}|(1-\epsilon)^n\leq\delta$; 2) prove $p(\hat R(\hat h)\geq\epsilon)\leq|\mathcal{H}|(1-\epsilon)^n$: $p(\mathcal{R}(\hat h)\geq\epsilon)\leq p(\exists h\in\mathcal{H}:\hat R(h)=0$ and $\mathcal{R}(h)\geq\epsilon)\leq\sum_{h\in\mathcal{H}}p(\hat R(h)=0|\mathcal{R}(h)\geq\epsilon)p(\mathcal{R}(h)\geq\epsilon)\leq\sum_hp(\hat R(h)=0|\mathcal{R}(h)\geq\epsilon)\leq\sum_h(1-\epsilon)^n$

VC Dimension: $VC(\mathcal{C})$ = max dimension n s.t. $\exists S\subseteq X,|S|=n$ and S can be shattered (any subset is bounded) by $\mathcal{C}$; e.g. VC(intervals) = 2.

Hoeffding's Theorem: $p(S_n-\mathbb{E}_XS_n\geq t)\geq exp(-\frac{2t^2}{\sum_i(b_i-a_i)^2})$; Proof: 1) $p(x\geq t) = p(exp(sX)\geq exp(st))\leq\frac{\mathbb{E}_X[exp(sX)]}{exp(st)}$; 2) $p(S_n-\mathbb{E}_XS_n\geq t)\leq e^{-st}\mathbb{E}_X[exp(s\sum_i(X_i-\mathbb{E}X_i))] = e^{-st}\prod_i\mathbb{E}_{X_i}[exp(s(X_i-\mathbb{E}X_i))]\leq e^{-st}\prod_iexp(s^2(b_i-a_i)^2/8)$, $s=\frac{4t}{\sum_i(b_i-a_i)^2}$

VC Inequality (distribution independent): For finite $\mathcal{C}$, $p(\mathcal{R}(\hat c_n^*) - inf_{c\in\mathcal{C}}(\mathcal{R}(c))\geq\epsilon)\leq p(\mathcal{R}(\hat c_n^*) - \hat{\mathcal{R}}(\hat c_n^*) + \hat{\mathcal{R}}(\hat c_n^*) - \mathcal{R}(c^*)\geq\epsilon)\leq p(2sup_c|\hat{\mathcal{R}}_n(c)-\mathcal{R}(c)|>\epsilon)\leq\sum_{c\in\mathcal{C}}p(|\hat{\mathcal{R}}_n(c)-\mathcal{R}(c)|>\epsilon)\leq2|\mathcal{C}|exp(-2n\epsilon^2)\Rightarrow R(c)$ exp. $\leq\hat R_n(c)$ emp. $+\sqrt{\frac{ln|C|-ln(\delta/2)}{2n}}$ var.