Lin Alg • $r_V \times c_V = s$ • $c_V \times r_V = M$ • $M \times c_V = c_V$ • $r_V \times M = r_V$ • $M \times M = M$ • $Au = \sum_{j=i} A^{(j)} u_j = c$ • $u^\top A = \sum_{j=i} A^{(i)} u_j = c^\top$ • $u^\top A u = \sum_i \sum_j x_i A_{ij} x_j$ • $AB = \sum_{j=i} A^{(j)} B^{(j)} = C$ • $(Au)_i = A^{(i)} u = \sum_j A_{ij} u_j$ • $(AB)_{ij} = A^{(i)} B^{(j)}$ • $(A^\top)^k = (A^k)^\top$ *Symm.* — $A^\top = A$ PD and PSD — PD: $x^\top A x > 0$, analog for PSD *Eigen.* $Aq = \lambda q$ *SVD* — $A = USV^\top$ with orth. U,V and sin. val. = sqrt. of the eigenval. of $A^\top A = AA^\top$ *Rules* — • $a^m \cdot a^n = a^{m+n}$ • $(ab)^n = a^n b^n$ • $\log(xy) = \log x + \log y$ *Polyn.* — $\sum_i a_i x^i$

Calc. *Rules* — • For $f(z,v), z=g(x), v=h(x)$: $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x}$ • $\partial a^\top x/\partial x = a$ • $\partial x^\top x/\partial x = 2x$ • $\partial Ax/\partial x = A^\top$ • $\partial x^\top Ax/\partial x = (A + A^\top)x$ *Constr. opt.* — *Lag.*: $\mathcal{L}(x,\lambda) = f(x) + \lambda g(x)$ For eq. constr.: $g(x)=0$ • Gradients are parallel at optimum: $\nabla_x f(x^*) = \lambda \times \nabla_x g(x^*)$ • To find x^* and λ^*: ∘ $\nabla_x \mathcal{L} = 0$ s.t $g(x)=0$ ∘ $\nabla_\lambda \mathcal{L}=0$ (constr.) • Optimum x^* and λ^* are saddle point of $\mathcal{L}$ For ineq. constr.: $g(x) \le 0$ • If x^* lies in $g(x)=0$, constr. is act. • Gradients are anti-parallel at optimum: $\nabla_x f(x^*) = -\lambda \times \nabla_x g(x^*)$ • To find x^* and λ^*: ∘ $\nabla_x \mathcal{L}=0$ s.t. *KKT conditions*: ■ $g(x)\le 0$ ■ $\lambda \ge 0$ ■ *Compl. slack. cond.*: $\lambda g(x)=0$, with $\lambda=0, g(x)<0$ (inact. constr.) and $\lambda>0, g(x)=0$ (act. constr.) ∘ $\nabla_\lambda \mathcal{L} = 0$ given compl. slack. cond. • Optimum x^* and λ^* are saddle point of $\mathcal{L}$ For multiple constr.: m ineq. constr. g and p eq. constr. h • Lag.: $\mathcal{L}(x,\lambda,\mu) = f(x) + \sum_{i=1}^m \mu^i g^i(x) + \sum_{j=1}^p \lambda^j h^j(x)$ • General sol. x^*, λ^*, μ^*: $\nabla_x \mathcal{L}=0$ s.t.: ∘ $\{g^i(x)\le 0\}_{i=1}^m$ and $\{h^j(x)=0\}_{j=1}^p$ ∘ $\{\mu^i \ge 0\}_{i=1}^m$ ∘ $\{\mu^i g^i(x)=0\}_{i=1}^m$ *Primal prob.*: • $\min_x[f(x)]$ s.t. constr. • Equivalent to $\min_x[\max_{\lambda,\mu} \mathcal{L}] = \min_x[f(x) + \max_{\lambda,\mu}[\sum_{i=1}^m \mu^i g^i(x) + \sum_{j=1}^p \lambda^j h^j(x)]]$ where term 2 is barrier function: ∘ $=0$ s.t. constr. being met ∘ $=\infty$ otherwise, insolvable *Dual prob.*: $\max_{\lambda,\mu}[\min_x \mathcal{L}]$ *Weak duality*: • Always holds • Given mini-max theorem, dual sol. is lower bound of primal sol.: $\min_x[\max_{\lambda,\mu}\mathcal{L}] = \min_x[f(x)] \ge \max_{\lambda,\mu}[\min_x \mathcal{L}]$ • $\min_x \mathcal{L}$ is an unconstr. opt. prob. • $\max_{\lambda,\mu}[\min_x \mathcal{L}]$ is a concave opt. prob. *Strong duality*: • Holds under *Slater's cond.* if there exists a sol. that strictly fulfills all ineq. constr. $\{g^i(x)<0\}_{i=1}^m$ i.e. lin. sep. • Dual sol. equals primal sol.: $\min_x[\max_{\lambda,\mu}\mathcal{L}] = \min_x[f(x)] = \max_{\lambda,\mu}[\min_x \mathcal{L}]$ • Solve $\min_x \mathcal{L}$ for primal var. x in terms of Lag. mult. λ,μ • Plug primal var. back into $\mathcal{L}$ and formulate addtl. constr. • Solve $\max_{\lambda,\mu}\mathcal{L}$ for Lag. mult. λ^*,μ^* • Specify x^* based on λ^*,μ^*

Prob. *Gaussian* — $X \sim N(\mu,\sigma^2)$, $1/\sigma\sqrt{2\pi}\exp(-(x-\mu)^2/2\sigma^2) = \ldots\exp(-x^2(1/2\sigma^2) + 2x(\mu/2\sigma^2) - (\mu^2/2\sigma^2))$ *Jensen's ineq.* — $\mathbb{E}(f(X)) \ge f(\mathbb{E}(X))$ *Total law* — $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$, $\mathbb{V}(X) = \mathbb{V}(\mathbb{E}[X|Y]) + \mathbb{E}[\mathbb{V}(X|Y)]$ *Var.* — • $\mathbb{V}[X] = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ • $\mathbb{V}[\sum_i x_i] = \sum_i \sum_j \mathrm{Cov}[x_i,x_j] = \sum_i \mathbb{V}[x_i] + \sum_{i\ne j}\mathrm{Cov}[x_i,x_j]$

Inf. Theory *Entropy* — $H = \mathbb{E}[-\log p(X)] = -\sum_x p(x)\log(p(x)) \ge 0$ *KL div.* — $KL = \mathbb{E}[\log(p(x)/q(x))] = \sum_x p(x)\log(p(x)/q(x)) = \sum_x p(x)\log(p(x)) - \sum_x p(x)\log(q(x)) = \sum_x p(x)\log(p(x)) - 1\times$const. if q(x) unif., $= -H(p) - \sum_x p(x)\log(q(x))$, ≥ 0 *Mut. inf.* — $I = \mathbb{E}[\log(p(x,y)/p(x)p(y))] = \sum_{x,y} p(x,y)\log(p(x,y)/p(x)p(y))$ *ELBO* — $\mathbb{E}[\log(p(x,z)/q(z))] = \sum_z q(z)\log(p(x,z)/q(z))$

ML Paradigms *Loss func.* — • Hinge loss: $\max(0,1-yz)$ • Exp. loss: e^{-yz} • Log. loss: $\log(1+e^{-yz})$

MLE • Max. log likelihood: $\hat\theta = \arg\max_\theta L(\theta) = (p(y_1,\ldots,y_n|x_i,\theta)) = (\prod_{i=1}^n p(y_i|x_i,\theta)) = (\sum_{i=1}^n \log(p(y_i|x_i,\theta)))$ • Discrete: ∘ $\hat\theta = \arg\max_\theta \prod_{j=1}^k p_j^{N_j} = \sum_{i=1}^n N_j \log(p_j) \propto \sum_{i=1}^n (N_j/n)\log(p_j) = \sum_{i=1}^n (N_j/n)(\log(p_j/(N_j/n)) + \log(N_j/n)) \propto \sum_{i=1}^n (N_j/n)\log(p_j/(N_j/n)) = \arg\min_\theta \sum_{i=1}^n (N_j/n)\log((N_j/n)/p_j) = \sum_{i=1}^n \tilde p_j \log(\tilde p_j/p_j)$, KL div. ∘ Solved via constr. opt. with strong duality s.t. $\sum_j p_j = 1$ ∘ $\theta_{MLE} = N_j/n$, min. the KL div. when $\tilde p_j = p_j$ • *Score*: ∘ Derivative of the log likelihood: $\Lambda = \frac{\partial}{\partial\theta}\log(p(y|x,\theta)) = \frac{\frac{\partial}{\partial\theta}p(y|x,\theta)}{p(y|x,\theta)}$ ∘ $\mathbb{E}(\Lambda)=0$ • Adv.: ∘ *Consistent*: $\hat\theta \to \theta$ as $n\to\infty$ ∘ *Asymp. normal*: $\frac{1}{\sqrt n}(\hat\theta - \theta)$ conv. to $N(0, J^{-1}(\theta) I(\theta) J^{-1}(\theta))$ where I is the Fisher inf. ∘ *Asymp. efficient*: $\hat\theta$ min. $\mathbb{E}[(\hat\theta-\theta)^2] \to \frac{1}{I_n(\theta)}$ as $n\to\infty$ where I is the Fisher inf. ∘ *Equivariant*: If $\hat\theta$ is MLE of θ, then $g(\hat\theta)$ is MLE of $g(\theta)$ • Proofs of Adv.: ∘ Asymp. normal: ■ Set score to 0: $\Lambda = \frac{\partial}{\partial\theta}\log(p(y|x,\theta))=0$ ■ With a Taylor exp.: $(\hat\theta-\theta)\sqrt n = \frac{1}{\sqrt n}\Lambda[-\frac{1}{n}\frac{\partial^2}{\partial\theta\partial\theta^\top}\sum_{i=1}^n p(y_i|x_i,\theta))]^{-1}$ ■ We set $J = [-\frac1n \frac{\partial^2}{\partial\theta\partial\theta^\top}\sum_{i=1}^n p(y_i|x_i,\theta)]$ ■ $\frac{1}{\sqrt n}\Lambda \sim N(0,I)$ ■ Then, $(\hat\theta - \theta)\sqrt n = J^{-1}\frac{1}{\sqrt n}\Lambda \sim J^{-1}N(0,I)$ ■ $\mathbb{V}(J^{-1}\frac{1}{\sqrt n}\Lambda) = \mathbb{E}[J^{-1}IJ^{-1}]$ because $\mathbb{E}(\Lambda)=0$ ■ Then: $(\hat\theta-\theta)\sqrt n \sim N(0, J^{-1}IJ^{-1})$ ∘ Equivariant: ■ Let $t = g(\theta)$ and $h = g^{-1}$ ■ Then, $\hat\theta = h(t) = h(g(\hat\theta))$ ■ For all t: $L(t) = \prod_i p(y_i|x_i,h(t))) = p(y_i|x_i,\theta) = L(\theta)$ ■ For all t: $L(t) = L(\theta)$ and $L(\hat t) = L(\hat\theta)$ • Equiv. to min. KL div. as $n\to\infty$: ∘ $\hat\theta_{MLE} = \arg\max_\theta \prod_{i=1}^n q(x_i|\theta) = \sum_{i=1}^n \log q(x_i|\theta)$ ∘ $\hat\theta_{KL} = \arg\min_\theta D_{KL}(\hat p(x)\|q(x|\theta))$ ∘ $D_{KL}(\hat p(x)\|q(x|\theta)) = \sum_{i=1}^n \hat p(x_i)\log\frac{\hat p(x_i)}{q(x_i|\theta)} = \frac1n\sum_{i=1}^n \log\frac{\hat p(x_i)}{q(x_i|\theta)} = \frac1n\sum_{i=1}^n\log\hat p(x_i) - \frac1n\sum_{i=1}^n\log q(x_i|\theta)$ ∘ Min. D_{KL} is equiv. to max.: $\frac1n\sum_{i=1}^n \log q(x_i|\theta)$

PAC • *Hyp. cl.* $\mathcal{H}$: Func. that can be expressed by the algo., VC-dim. det. req. samples • *Concept cl.* C: Func. that repr. true mappings from input to output • *Learning algo.* $\mathcal{A}$: Outputs $\hat c$ from $\mathcal{H}$ PAC: • $P(\mathcal{R}(\hat c)\le\epsilon)\ge 1-\delta$, i.e. $P(\mathcal{R}(\hat c)\le\epsilon)\ge 1-\delta$ where: ∘ ϵ: Error tolerance (how much $\hat c$ deviates from c) ∘ δ: Confidence (how likely $\hat c$ generalizes well) • If algo. $\mathcal{A}$ runs in time polyn. in $1/\epsilon$ and $1/\delta$, then C is *efficiently PAC-learnable* 1) C is finite, $\mathcal{H}=C$: • For single inst. $P(\mathcal{R}(\hat c)\ne 0) = \epsilon$ and $P(\mathcal{R}(\hat c)=0) = (1-\epsilon)$ • For all inst.: $P(\mathcal{R}(\hat c)=0) = (1-\epsilon)^n$ • Acc. to Hoeffding's ineq.: $(1-\epsilon)^n \le \exp(-n\epsilon)$ • Then: $P(|\mathcal{R}(c)-\hat{\mathcal{R}}(c)|>\epsilon) = P(\exists c\in C: \mathcal{R}(\hat c)=0$ and $\mathcal{R}(c)>\epsilon) = (1-\epsilon)^n \le \exp(-n\epsilon)$ • With union bound: $P(\exists c\in C: \mathcal{R}(\hat c)=0) \le |C|(1-\epsilon)^n$ • $P(\exists c\in C: \mathcal{R}(\hat c)=0$ and $\mathcal{R}(c)>\epsilon) = |C|(1-\epsilon)^n \le |C|\exp(-n\epsilon)$ • Since we want $P(\mathcal{R}(c)>\epsilon)\le\delta$, we can set up: $|C|\exp(-n\epsilon)\le\delta$ • Then: $\frac{1}{\epsilon}[\log(|C|) + \log(\frac1\delta)] \le n$ E.g.: • $c(x) = \begin{cases}0 & \text{if } x\in[a,b]\\ 1 & \text{otherwise}\end{cases}$ • Let $[X^n_{min}, X^n_{max}]$ be the learned interval where inst. are cl. as 1 • Let $P(a \le X_i < a_\epsilon) = P(b_\epsilon < X_i \le b) = \epsilon/2$, i.e. this delta repr. a normal range of error • Then: Event $X^n_{max} \le b_\epsilon$ means none of the X_i fall into $[b_\epsilon, b]$ • So: $P(X^n_{max} \le b_\epsilon) = P(X_i \notin [b_\epsilon, b]$ for $\forall i) = \prod_{i=1}^n P(X_i \notin [b_\epsilon, b]) = \prod_{i=1}^n 1-\epsilon/2 = (1-\epsilon/2)^n$ • We want: $P(R(c)>\epsilon)\le\delta$ • By union bound: $P(R(\hat c)>\epsilon) = P(X^n_{max}\le b_\epsilon) + P(X^n_{min}\ge a_\epsilon) = 2(1-\epsilon/2)^n$ • We can re-form $2(1-\epsilon/2)^n \le 2\exp(-n\epsilon/2)\le\delta$ due to Hoeffding's ineq. and then to $\exp(n\epsilon/2)\ge\frac2\delta$ to $n \ge \frac2\epsilon\log(\frac2\delta)$ • C is efficiently PAC learnable since $\mathcal{A}$ runs in time polyn. in $1/\epsilon$ and $1/\delta$ 2) C is finite, $\mathcal{H}\ne C$ 3) C is infinite with finite VC-dim. N and unif. conv. $P(\mathcal{R}(\hat c^*) - \inf_{c\in C}\mathcal{R}(c) > \epsilon) \le P(\sup_{c\in C}|\hat{\mathcal{R}}(c) - \mathcal{R}(c)| > \frac\epsilon2) \le \sum_{c\in C} P(|\hat{\mathcal{R}}(c)-\mathcal{R}(c)|>\epsilon) \le 2|C|\exp(-n\frac{\epsilon}{2}^2) = \delta$

MAP • Max. post.: $\hat\theta = \arg\max_\theta(p(\theta|X)) \propto p(\theta|X)p(X)$ • Discrete: ∘ Make $\theta_{MLE} = N_j/n$ more robust with prior $p(\theta) \propto \prod_{j=1}^k p_j^v$ with param. $0<v\le 1$ ∘ $\hat\theta = \arg\max_\theta(p(\theta|y)) = (p(y|\theta)p(\theta)) = \prod_{j=1}^k p_j^{N_j}\prod_{j=1}^k p_j^v = \prod_{j=1}^k p_j^{N_j+v} = \sum_{j=1}^k (N_j+v)\log(p_j) \propto \sum_{j=1}^k \frac{N_j+v}{n+kv}\log(\frac{p_j}{(N_j+v)/(n+kv)}) = \arg\min_\theta \sum_{j=1}^k \frac{N_j+v}{n+kv}\log(\frac{(N_j+v)/(n+kv)}{p_j}) = \sum_{j=1}^k \tilde p_j\log(\tilde p_j/p_j)$, KL div. • This can be solved using constr. opt. with strong duality s.t. $\sum_j p_j = 1$ ∘ $\theta_{MAP} = (N_j+v)/(n+kv)$, min. the KL div.

Model Taxonomy

Active Learning • Data $D_{n-1}(x_i,y_i)_{i<n}$ • Find the next x_n that gives most info, i.e. max. the cond. mutual inf. between y_x and f: $IG[f_X; y_X|D_{n-1}] = H(y_X|D_{n-1}) - H(y_X|f_X, D_{n-1})$ where $H(y_X|f_X,D_{n-1})$ is the uncert. about y_X after labeling x_n • Pick $x_n = \arg\max_{x\in S} IG$ • f is a GP: ∘ $f\sim\mathcal{GP}(\mu,k)$ ∘ $f = (f_{X_1}, f_{X_2},\ldots)\sim N(\mu,\Sigma)$ with $\mu_i = \mu(x_i)$ and $\Sigma_{ij} = k(x_i,x_j)$ • Then: $IG = \frac12\log(\frac{\mathbb{V}(y_X|D_{n-1})}{\mathbb{V}(y_X|f_X,D_{n-1})})$ Proof: ∘ Gaussian entropy of $a|B$: $H(a|B) = -\int p(a|B)\log(p(a|B))da = -\mathbb{E}[\log N(\mu,\sigma)] = \frac12\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}[(x-\mu)^2] = \frac12\log(2\pi\sigma^2) + \frac12\times 1 = \frac12\log(2\pi\sigma^2) + \frac12\log(e) = \frac12\log(2\pi e\sigma^2)$ ∘ Plugging in: $H(y_X|D_{n-1}) - H(y_X|f_X,D_{n-1}) = \frac12\log(2\pi e\mathbb{V}(y_X|D_{n-1})) - \frac12\log(2\pi e\mathbb{V}(y_X|f_X,D_{n-1}))$

Ensemble Methods *Boost.* — Red. bias • Avg. unbiased. Proof: Bias $= \mathbb{E}[\hat f(x)] - \mathbb{E}[y|x] = \frac1B\sum_{i=1}^B\mathbb{E}[\hat f_i(x)] - \mathbb{E}[y|x] = \frac1B\sum_{i=1}^B \text{bias}(\hat f_i(x))$ • Var. red. by fact. $\frac1B$, if est. have similar var. ($\approx\sigma^2$) and no cov. Proof: ∘ Variance $= \mathbb{E}[\hat f(x) - \mathbb{E}[\hat f(x)]]^2 = \mathbb{E}[\frac1B\sum_{i=1}^B \hat f_i(x) - \frac1B\sum_{i=1}^B\mathbb{E}[\hat f_i(x)]]^2 = \mathbb{E}[\frac1B\sum_{i=1}^B(\hat f_i(x) - \mathbb{E}[\hat f_i(x)])]^2 = \frac{1}{B^2}\sum_{i=1}^B\mathbb{V}[\hat f_i(x)] + \frac{1}{B^2}\sum\sum_{i\ne j=1}^B\mathrm{Cov}[\hat f_i(x),\hat f_j(x)] = \frac{B}{B^2}\mathbb{V}[\hat f_i(x)] + \frac{B(B-1)}{B^2}\mathrm{Cov}[\hat f_i(x),\hat f_j(x)]$ *AdaBoost* — Algorithm: 1) Instance weight: $w_i = 1/n$ 2) Train $\hat c^{(1)}$ with output $\hat y^{(1)}$ 3) Calc. error of j^{th} clf.: $r^{(j)} = (\sum_{i=1}^n w_i I_{\hat y \ne y_i})/(\sum_{i=1}^n w_i)$ 4) Calc. clf. weight (high if has low error): $\alpha^{(j)} = \eta\log([1-r^{(j)}]/[r^{(j)}])$ 5) Update instance weights: For $i=1,\ldots,n$: $w_i = w_i e^{\alpha^{(j)}}$ if $\hat y_i^{(j)} \ne y_i$ else w_i 6) Norm. instance weights: $w_i = \frac{w_i}{\sum_{i=1}^n w_i}$ 7) Train next clf. 8) ... 9) Pred.: $\hat y(x) = \arg\max_k \sum_{j=1}^B \alpha^j I_{\{\hat c^{(j)}(x)=k\}}$

Model Evaluation

Estimator • Exp. loss $\mathbb{E}[y - x^\top\beta]^2$ • Emp. loss $\frac1n\sum_i(y_i - x_i^\top\beta)^2$ • Consistency: $\hat\theta\to\theta$ as $n\to\infty$ • Bias: $\mathbb{E}(\hat\theta) - \theta$ ∘ Unbiased: $\mathbb{E}(\hat\theta) = \theta$ ∘ Asymp. unbia.: $\mathbb{E}[(\hat\theta-\theta)^2] = 0$ as $n\to\infty$ ∘ Asymp. efficient: $\mathbb{E}[(\hat\theta-\theta)^2] = I$ as $n\to\infty$ where I is Fisher inf. *Fisher inf.* — • $I = \mathbb{E}[(\Lambda)^2] = \mathbb{E}[(\frac{\partial}{\partial\theta}\log(p(x|\theta)))^2] = \mathbb{V}(\frac{\partial\log(p(x|\theta))}{\partial\theta})$ where Λ is the score and $\mathbb{E}(\Lambda)=0$ • For n iid samples: $I_n = n\times I$ Proof: ∘ $\mathbb{E}[(\frac{\partial}{\partial\theta}\log(p(x^{(1)},\ldots,x^{(n)}|\theta)))^2] = \mathbb{E}[(\frac{\partial}{\partial\theta}\log(\prod_{i=1}^n p(x^i|\theta)))^2] = \mathbb{E}[(\sum_{i=1}^n\frac{\partial}{\partial\theta}\log(p(x^i|\theta)))^2] = \mathbb{E}[\sum_{i=1}^n(\frac{\partial}{\partial\theta}\log(p(x^i|\theta))^2 + \sum_{i\ne j}(\frac{\partial}{\partial\theta}\log(p(x^i|\theta))\times\frac{\partial}{\partial\theta}\log(p(x^{(j)}|\theta)))]$ ∘ $= \mathbb{E}[\sum_{i=1}^n(\frac{\partial}{\partial\theta}\log(p(x^i|\theta))^2] + \sum_{i\ne j}\mathbb{E}[(\frac{\partial}{\partial\theta}\log(p(x^i|\theta))\times\mathbb{E}[\frac{\partial}{\partial\theta}\log(p(x^{(j)}|\theta))]$ since cross-terms are independent ∘ $= \mathbb{E}[\sum_{i=1}^n(\frac{\partial}{\partial\theta}\log(p(x^i|\theta))^2]$ since expected value of score is 0 ∘ $= \sum_{i=1}^n\mathbb{E}[(\frac{\partial}{\partial\theta}\log(p(x^i|\theta))^2] = n\times\mathbb{E}[(\frac{\partial}{\partial\theta}\log(p(x^i|\theta))^2]$ • Small var., large Fisher inf., small Rao-Cramer bound, small MSE *Rao-Cramer bound* — • No asymptotically unbia. est. • For each unbia. est.: $\mathbb{E}[(\hat\theta-\theta)^2] \ge \frac1I$ where I is the Fisher inf. • For all est.: $\frac{(\frac{\partial}{\partial\theta}\text{bias}+1)^2}{I} + \text{bias}^2 \le \mathbb{E}[(\hat\theta-\theta)^2]$ Proof: ∘ Given Cauchy Schwarz ineq.: $\mathbb{E}[(\Lambda-\mathbb{E}((\Lambda))(\hat\theta-\mathbb{E}(\hat\theta))]^2 \le \mathbb{E}[(\Lambda-\mathbb{E}(\Lambda))^2]\mathbb{E}[(\hat\theta-\mathbb{E}(\hat\theta))^2]$ where Λ is the score ∘ We know that $\mathbb{E}(\Lambda)=0$ ∘ LHS: ■ Since $\mathbb{E}(\Lambda)=0$, we can simplify to $\mathbb{E}[\Lambda(\hat\theta-\mathbb{E}(\hat\theta))] = \mathbb{E}[\Lambda\hat\theta] - \mathbb{E}[\Lambda]\mathbb{E}[\hat\theta] = \mathbb{E}[\Lambda\hat\theta] - 0$ ■ This can be developed to: $\mathbb{E}[\Lambda\hat\theta] = \int p(x|\theta)\frac{\frac{\partial}{\partial\theta}p(x|\theta)}{p(x|\theta)}\hat\theta dx = \frac{\partial}{\partial\theta}(\int p(x|\theta)\hat\theta dx - \theta) + 1$ where the last part $(-\theta)+1$ can be added, because $\frac{\partial}{\partial\theta}-\theta = -1$ and we compensate this with +1 ■ Equal to the derivative of the bias + 1: $\frac{\partial}{\partial\theta}(\int p(x|\theta)\hat\theta dx - \theta) + 1 = \frac{\partial}{\partial\theta}(\mathbb{E}[\hat\theta]-\theta) + 1 = \frac{\partial}{\partial\theta}\text{bias}+1$ ∘ RHS: Since $\mathbb{E}(\Lambda)=0$, first term is $\mathbb{E}(\Lambda^2) = I$ where I is the Fisher inf. ∘ Then: $(\frac{\partial}{\partial\theta}\text{bias}+1)^2 \le I\times\mathbb{E}[(\hat\theta-\mathbb{E}(\hat\theta))^2] = I\times\mathbb{E}[(\hat\theta-\theta-\mathbb{E}(\hat\theta)+\theta)^2] = \ldots = I\times\mathbb{E}[(\hat\theta-\theta)^2] - \text{bias}^2$ ∘ Then: $\frac{(\frac{\partial}{\partial\theta}\text{bias}+1)^2}{I} + \text{bias}^2 \le \mathbb{E}[(\hat\theta-\theta)^2]$

Bias Variance Tradeoff • MSE $\mathbb{E}[(\hat f(X) - y)^2]$: $(\mathbb{E}[\hat f(X)] - f(X))^2 + \mathbb{V}(\hat f(X)) + \mathbb{E}[\epsilon^2] =$ bias^2+variance+irred. error Proof: ∘ $y = f(X) + \epsilon$ ∘ $\mathbb{E}[(\hat f(X)-y)^2] = \mathbb{E}[(\hat f(X) - \mathbb{E}[\hat f(X)] + \mathbb{E}[\hat f(X)] - f(X) + \epsilon)^2] = \mathbb{E}[(\hat f(X) - \mathbb{E}[\hat f(X)])^2] + \mathbb{E}[(\mathbb{E}[\hat f(X)] - f(X))^2] + \mathbb{E}[\epsilon^2] - 2\mathbb{E}[(\hat f(X) - \mathbb{E}[\hat f(X)])(\mathbb{E}[\hat f(X)] - f(X) + \epsilon)]$ ∘ Term 3 $\mathbb{E}[\epsilon^2]$ is the variance of y: $= \mathbb{E}[\epsilon^2] - \mathbb{E}[\epsilon]^2 = \mathbb{V}(y) = \sigma^2$ ∘ Term 4 $2\mathbb{E}[(\hat f(X) - \mathbb{E}[\hat f(X)])(\mathbb{E}[\hat f(X)] - f(X) + \epsilon)]$ equals 0: ■ $2\mathbb{E}[(\hat f(X) - \mathbb{E}[\hat f(X)])(\mathbb{E}[\hat f(X)] - f(X) + \epsilon)] = 2(\mathbb{E}[\hat f(X)] - f(X) + \epsilon)\mathbb{E}[(\hat f(X) - \mathbb{E}[\hat f(X)])]$ because $(\mathbb{E}[\hat f(X)] - f(X) + \epsilon)$ is deterministic ■ In last equation, second term equals 0, so whole equation is 0 ∘ Then, we are left with: variance+bias^2+irred. error

Approx. Gen. Loss *Bagg.* — Red. var. *Bootstr.* — • B samples with repl. $\mathcal{Z}$: $\mathcal{Z}^{*b}$ • Prob. that x_i is in $\mathcal{Z}$ 1 time: $\binom{n}{1}(\frac1n)^1(\frac{n-1}{n})^{n-1} = (\frac{n-1}{n})^{n-1}$ • Est. $S(\mathcal{Z}^{*b})$ for each sample • For each est.: ∘ $\bar S = \frac1B\sum_b S(\mathcal{Z}^{*b})$ ∘ $\sigma^2(S) = \frac{1}{B-1}\sum_b(S(\mathcal{Z}^{*b}) - \bar S)^2$ • OOB loss: $\hat{\mathcal{R}}^{bs} = \frac1n\sum_{i=1}^n\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}}LO(y_i - \hat f^{*b}(x_i))$ where C^{-i} contains all bootstr. ind. b so that $\mathcal{Z}^{*b}$ does not contain (x_i,y_i) • Emp. loss: $\hat{\mathcal{R}}(\mathcal{A}) = \frac1B\sum_{b=1}^B\frac1n\sum_{i=1}^n LO(y_i - \hat f^{*b}(x_i))$ • Emp. loss uses tr. dat., too optimistic, true loss for 1/3 of dat., close to 0 for other 2/3. Correction: ∘ Prob. that (x_i,y_i) is not in sample $\mathcal{Z}^{*b}$ of size n is $(1-\frac1n)^n = \frac1e$ as $n\to\infty \approx \frac13$ ∘ Prob. that (x_i,y_i) is in sample is $1-\frac1e$ as $n\to\infty\approx\frac23$ ∘ Then: $\hat{\mathcal{R}}^{(0.632)} = 0.368\hat{\mathcal{R}}(\mathcal{A}) + 0.632\hat{\mathcal{R}}^{bs}$ *Grad. desc.* — $\theta_{t-1} - \eta\nabla\mathcal{L}(\theta_{t-1})$

K-Means Clustering

Form. • Clust. with centroid $\mu^{[j]}\in\mathbb{R}^m$ • k ind. var. if inst. i is assigned to clust. j: $\{p^{i[j]}\}_{j=1}^k \in [0,1]$

Opt. *Parameters* — Find centroids $\mu^{[j]}$ *Objective function* — Min. dist.: $j^* = \arg\min_j\sum_i\sum_j p^{i[j]}\|x^i - \mu^{[j]}\|^2$ *Optimization* — Lloyd's algo.: 1) Rand. init. each $\mu^{[j]}$ 2) E-Step: Re-assign inst while keeping all centroids fixed, i.e. minimize Θ wrt. $p^{i[j]}$ 3) M-Step: Re-compute centroids while keeping all inst. assignm. fixed, i.e. minimize Θ wrt. to $\mu^{[j]}$ • $\nabla_{\mu^{[j]}}\Theta = 2\sum_{i=1}^n p^{i[j]}(\mu^{[j]} - x^i) = 0$ • Then, $\mu^{[j]} = \frac{\sum_{i=1}^n p^{i[j]}x^i}{\sum_{i=1}^n p^{i[j]}}$ • Strictly convex 4) Repeat Proof of conv. $\Theta_t \le \Theta_{t-1}$: • $\Theta_t = \sum_{i=1}^n\sum_{j=1}^k p_t^{i[j]}\|x^i - \mu_t^{[j]}\|^2$ • In M-Step: $\Theta_t \le \sum_{i=1}^n\sum_{j=1}^k p_t^{i[j]}\|x^i - \mu_{t-1}^{[j]}\|^2$ • In E-Step: $\sum_{i=1}^n\sum_{j=1}^k p_t^{i[j]}\|x^i - \mu_{t-1}^{[j]}\|^2 \le \sum_{i=1}^n\sum_{j=1}^k p_{t-1}^{i[j]}\|x^i - \mu_{t-1}^{[j]}\|^2 = \Theta_{t-1}$

NP Bayes

CRP • Cust. are obs. $\theta^{[i]}$ • Tab. are clust. θ_k • Let $\mathcal{P}$ be the current seating • Prob. that cust. $n+1$ joins T is: $P(\text{cust. } n+1 \text{ joins } T|\mathcal{P}) = \begin{cases}\frac{|T|}{\alpha+n} & \text{if } T \text{ is in } \mathcal{P}\\ \frac{\alpha}{\alpha+n} & \text{otherwise}\end{cases}$ • Prob. of seating: $P(\mathcal{P}) = P(\text{cust. 1 joins T})\times P(\text{what cust. 2 does (join or start new tab.)}), \ldots = \frac{\alpha^{|\mathcal{P}|}}{\alpha^{(n)}}\prod_{T\in\mathcal{P}}(|T|-1)!$ • Expected number of tab.: $\mathbb{E}(|\mathcal{P}|) = O(\alpha\log(N))$ CRP for cont. dist.: • $p(x_k) = \frac{\alpha}{\alpha+k-1}$ and $\sum_{k=1}^n p(x_k) = \sum_{k=1}^n\frac{\alpha}{\alpha+k-1} = S(n)$ • Approx. sum $S(n)$ with int. $I(n)$: $I(n) = \int_1^{n+1}\frac{\alpha}{\alpha+x-1}dx$ • Int. is bounded above by the sum: $I(n)\le S(n)$ • Int. is bounded below by: $I(n)\ge\sum_{k=2}^{n+1}\frac{\alpha}{\alpha+k-1}$ • Therefore: $I(n)\ge S(n) - \frac\alpha\alpha + \frac{\alpha}{\alpha+n} = S'(n)$ • Then: $S(n) - 1 + \frac{\alpha}{\alpha+n} \le I(n) \le S(n)$ and equiv.: $I(n) + 1 - \frac{\alpha}{\alpha+n} \ge S(n) \ge I(n)$ • Comp. the int. $I(n)$: $I(n) = \int_1^{n+1}\frac{\alpha}{\alpha+x-1}dx = F(n+1) - F(1) = \alpha(\ln(\alpha+n) - \ln(\alpha))$ • As $n\to\infty$, $I(n)\to\alpha(\ln(n))$. Thus, as $n\to\infty$, $S(n)\to\alpha(\ln(n))$

DP mixture model • Let: ∘ $\Theta=\mathbb{R}$ be space of parameters ∘ $H = N(\mu_0,\sigma_0)$ be base measure on Θ • DP mixture model: ∘ Clust. prob.: $\pi = (\pi^{[1]},\ldots)\sim GEM(\alpha)$ ∘ Clust. centers: $\mu_k \sim N(\mu_0,\sigma_0)$ ∘ Assignm. of datapoints to clust.: $z^{(i)}\sim\mathrm{Cat}(\pi)$ ∘ Coordinates of datapoints: $x^{(i)}\sim N(\mu^{[z^{(i)}]}, \sigma^{[z^{(i)}]})$ • Fitting the model: ∘ Prior: Prob. of clust. assignm., based on clust. size ∘ Likelihood: Prob. of datapoint, given clust. center ∘ *Gibbs sampling*: ■ $P(z^{(i)}=k|z^{(-i)},x,\alpha,\mu) \propto P(z^{(i)}=k|z^{(-i)})\,P(x|z^{(i)}=k,z^{(-i)})$ due to Bayes ■ $\propto P(z^{(i)}=k|z^{(-i)})\,P(x^{(i)}|z^{(i)}=k,z^{(-i)},x^{(-i)})\,P(x^{(-i)}|z^{(i)}=k,z^{(-i)})$ due to prod. rule ■ $\propto P(z^{(i)}=k|z^{(-i)})\,P(x^{(i)}|z^{(i)}=k,z^{(-i)},x^{(-i)})$ because in last term $x^{(-i)}\perp z^{(i)}|z^{(-i)}$ by d-sep., so this term is const. wrt. to k ■ Then: $P(z^{(i)}=k|z^{(-i)},\alpha)\,P(x^{(i)}|z^{(i)}=k,z^{(-i)},x^{(-i)},\mu)$ = prior × likelihood, with interchangeable $z^{(1)},\ldots,z^{(k)}$ ■ Prior: From CRP: $P(z^{(i)}=k|z^{(-i)},\alpha) = \begin{cases}\frac{N_{k,-i}}{\alpha+N-1} & \text{if } k \text{ is an existing clust.}\\ \frac{\alpha}{\alpha+N-1} & \text{otherwise}\end{cases}$ ■ Likelihood: Only consider points assign. to clust. k: $P(x^{(i)}|z^{(i)}=k,z^{(-i)},x^{(-i)},\mu) = \begin{cases}P(x^{(i)}|x^{(-i,k)},\mu) = \frac{P(x^{(i)},x^{(-i,k)}|\mu)}{P(x^{(-i,k)}|\mu)} & \text{for ext. } k\\ P(x^{(i)}|\mu) & \text{otherwise}\end{cases}$ ■ Thus, Gibbs sampler: $P(z^{(i)}=k|z^{(-i)},x,\alpha,\mu) = \begin{cases}\frac{N_{k,-i}}{\alpha+N-1}\times P(x^{(i)}|x^{(-i,k)},\mu) & \text{for ext. } k\\ \frac{\alpha}{\alpha+N-1}\times P(x^{(i)}|\mu) & \text{otherwise}\end{cases}$

PCA • Axes = PCs • Axis = eigenvector $u^{[j]}$ with *loadings* $[u_1^{[j]},\ldots,u_m^{[j]}]$ • Var. along each axis = eigenvalue

Max. Var. Approach – Form. • Proj. data $\{x^i\}_{i=1}^n\in\mathbb{R}^m$ onto space $\mathbb{R}^d$ spanned by orthon. basis $\{u^{[j]}\}_{j=1}^d\in\mathbb{R}^m$ where $d<<m$ • Var.: $\frac1n\sum_{i=1}^n(u^{[j]\top}x^i - u^{[j]\top}\bar x)^2 = u^{[j]\top}Su^{[j]}$ where $S = \frac1n\sum_{i=1}^n(x^i - \bar x)(x^i - \bar x)^\top = \frac1n X^\top X$ is the covar. matrix of the data in the orthog. compl. to the subsp. spanned by the first $j-1$ PCs, i.e. $X_{j-1} = \{x^i - \text{projection}_{u_{\le j-1}}\} = \{x^i - \sum_{l=1}^{j-1}(u^{[l]}\cdot x^i)\cdot u^{[l]}\}$

Opt. *Parameters* — Find PCs $\{u^{[j]}\}_{j=1}^d$ *Objective function* — • For the j^{th} PC, maximize var. $\sum_{j=1}^d u^{[j]\top}Su^{[j]}$ s.t. orthon. $\{u^{[j]}\}_{j=1}^d$ • Lag. for $u^{[1]}$: $\mathcal{L} = u^{[1]\top}Su^{[1]} - \lambda^{[1]}(u^{[1]}\cdot u^{[1]} - 1)$ • Lag. for $u^{[2]}$ addtl. has orthog. constr. *Optimization* — For $u^{[1]}$: • $\nabla_{u^{[1]}}\mathcal{L} = 2Su^{[1]} - 2\lambda^{[1]}u^{[1]} = 0$ • $\Rightarrow Su^{[1]} = \lambda^{[1]}u^{[1]}$ • We see $u^{[1]}$ is the eigenvector of S and $\lambda^{[1]}$ is the eigenvalue • We see that the var. of the proj. data is equal to $\lambda^{[1]}$: $u^{[1]\top}Su^{[1]} = u^{[1]\top}\lambda^{[1]}u^{[1]} = \lambda^{[1]}u^{[1]\top}u^{[1]} = \lambda^{[1]}\times 1$ Total proj. var. of $S = \sum_{j=1}^d\lambda^{[j]}$

SVD Approach – Form. • Proj. data $\{x^i\}_{i=1}^n\in\mathbb{R}^m$ onto space $\mathbb{R}^d$ spanned by orthon. basis $\{u^{[j]}\}_{j=1}^d\in\mathbb{R}^m$ (subset of $\{u^{[j]}\}_{j=1}^m\in\mathbb{R}^m$) where $d\ll m$ • Original datapoint: $x^i = \sum_{j=1}^m(x^i\cdot u^{[j]})u^{[j]}$ • Proj. datapoint: $\tilde x^i = \sum_{j=1}^d\alpha_{ij}u^{[j]} + \sum_{j=d+1}^m\gamma_j u^{[j]}$ where α_{ij} is spec. to inst. and γ_j is generic

Opt. *Objective function* — • Reconstr. error: $J = \frac1n\sum_{i=1}^n\|x^i - \tilde x^i\|^2$ • $\alpha_{ij} = x^i\cdot u^{[j]}$ and $\gamma_j = \bar x\cdot u^{[j]}$ min. the reconstr. error • If $d=m-1$, $J = u^{[D]\top}Su^{[D]}$: ∘ For $d=m-1$, $\tilde x^i = \sum_{j=1}^d\alpha_{ij}u^{[j]}$ ∘ Then we get: $J = \frac1n\sum_{i=1}^n\|x^i - \tilde x^i\|^2 = \frac1n\sum_{i=1}^n\|\sum_{j=1}^m(x^i\cdot u^{[j]})u^{[j]} - \sum_{j=1}^{m-1}\alpha_{ij}u^{[j]}\|^2 = \frac1n\sum_{i=1}^n\|\sum_{j=1}^m(x^i\cdot u^{[j]})u^{[j]} - \sum_{j=1}^{m-1}(x^i\cdot u^{[j]})u^{[j]}\|^2 = \frac1n\sum_{i=1}^n\|(x^i\cdot u^{[D]})u^{[D]}\|^2$ ∘ Using orthon. of $u^{[D]}$: $\|(x^i\cdot u^{[D]})u^{[D]}\|^2 = (x^i\cdot u^{[D]})^2$ ∘ Thus: $J = \frac1n\sum_{i=1}^n(x^i\cdot u^{[D]})^2$ ∘ Center data: $J = \frac1n\sum_{i=1}^n((x^i - \bar x)\cdot u^{[D]})^2$ ∘ Expanding: $J = \frac1n\sum_{i=1}^n((x^i - \bar x)\cdot u^{[D]})^\top((x^i - \bar x)\cdot u^{[D]}) = u^{[D]\top}(x^i - \bar x)(x^i - \bar x)^\top u^{[D]} = u^{[D]\top}Su^{[D]}$ where S is covar. matrix and $u^{[D]}$ is its eigenvec. with the smallest eigenval., due to SVD of $S = V\Lambda V^\top$

GMM

Form. • Inst. $\{x^i\}_{i=1}^n$ • Inst. with latent cluster assignm.: $z^i\in\{1,\ldots,k\}$ • Prob. that cluster assign. to inst. i is cluster j: $\pi^{[j]} = p(z^i = j)$ • $x^i|z^i = j\sim N(\mu^{[j]},\Sigma^{[j]})$ • Marg. dist. of each inst.: $p(x^i) = \sum_{j=1}^k p(x^i|z^i)p(z^i) = \sum_{j=1}^k N(x^i|\mu^{[j]},\Sigma^{[j]})\times\pi^{[j]}$ • GMM, characterized by $\theta = \{\mu^{[j]},\Sigma^{[j]},\pi^{[j]}\}_{j=1}^k$

Opt. *Objective function* — In multi-class case: • Max. log likelihood $L = \sum_{i=1}^n\log p(x^i) = \sum_{i=1}^n\log(\sum_{j=1}^k p(x^i|z^i)p(z^i)) = \sum_{i=1}^n\log(\sum_{j=1}^k N(x^i|\mu^{[j]},\Sigma^{[j]})\times\pi^{[j]})$ subject to $\sum_{j=1}^k\pi^{[j]} = 1$ and $\Sigma^{[j]}\succ 0$ • Challenge: Sum in the log • Solution: EM algo. • Assume we know cluster assignm. z^i for each inst. • For each inst., let q_{z^i} be some prob. dist. over z^i • Then: $L = \sum_{i=1}^n\log\sum_{z^i}q(z^i)\frac{p(x^i,z^i)}{q(z^i)}$ • According to Jensen's ineq.: $L = \sum_{i=1}^n\log\sum_{z^i}q(z^i)\frac{p(x^i,z^i)}{q(z^i)} \ge \sum_{i=1}^n\sum_{z^i}q(z^i)\log\frac{p(x^i,z^i)}{q(z^i)}$ • Derived a lower bound for L • Tighten this bound by selecting: $q_{z^i} = p(z^i|x^i)$ ∘ Proof 1: $RHS = \sum_{i=1}^n\sum_{z^i}p(z^i|x^i)\log\frac{p(x^i,z^i)}{p(z^i|x^i)} = \sum_{i=1}^n\sum_{z^i}p(z^i|x^i)\log\frac{p(x^i|z^i)p(z^i|x^i)}{p(z^i|x^i)} = \sum_{i=1}^n\sum_{z^i}p(z^i|x^i)\log p(x^i) = \sum_{i=1}^n 1\times\log p(x^i) = L$ ∘ Proof 2: $L = \mathbb{E}[\log(p(x^i))] = \mathbb{E}[\log(\frac{p(x^i,z^i)}{p(z^i|x^i)}\frac{q(z^i)}{q(z^i)})] = \mathbb{E}[\log(\frac{p(x^i,z^i)}{q(z^i)})] + \mathbb{E}[\log(\frac{q(z^i)}{p(z^i|x^i)})]$ = M+E (KL div.) • Specify: $q_{z^i} = p(z^i|x^i) = \frac{p(x^i|z^i)p(z^i)}{\sum_{j=1}^k p(x^i|z^i)p(z^i)} = \frac{N(x^i|\mu^{[j]},\Sigma^{[j]})\times\pi^{[j]}}{\sum_{j=1}^k N(x^i|\mu^{[j]},\Sigma^{[j]})\times\pi^{[j]}} = \gamma^{i[j]}$ In binary case: • For each inst., let $M_{z^i} = 1$ if cluster $z^i = j$ has generated x^i • Then, joint likelihood: $\prod_{i=1}^n\prod_{z^i}(p(x^i|z^i)p(z^i))^{M_{z^i}} = \prod_{i=1}^n\prod_{z^i}(p(x^i|z^i)\pi^{[j]})^{M_{z^i}}$ • Log likelihood: $\sum_{i=1}^n\sum_{z^i}M_{z^i}\log(p(x^i|z^i)\pi^{[j]})$ • Expec. over latent var. M, given θ: $\mathbb{E}[\sum_{i=1}^n\sum_{z^i}M_{z^i}\log(p(x^i|z^i)\pi^{[j]})] = \sum_{i=1}^n\sum_{z^i}\mathbb{E}[M_{z^i}|X]\log(p(x^i|z^i)\pi^{[j]})$ • Then: $\mathbb{E}[M_{z^i}|X] = P(M_{z^i}=1|x^i)\times 1 + P(M_{z^i}=0|x^i)\times 0 = P(M_{z^i}=1|x^i) = P(z^i=j|x^i) = [P(x^i|z^i=j)P(z^i=j)]/[P(x^i)]$

Optimization — *EM algo.* 1) Rand. init. $\theta^{(t)} = \{\mu^{[j](t)},\Sigma^{[j](t)},\pi^{[j](t)}\}_{j=1}^k$ 2) *E-step*: Min. E, by comp. $q(z^i)$ given x^i and $\theta^{(t)}$ 3) *M-step*: Max. M, by upd. $\theta^{(t)}$ keeping $q(z^i)$ fixed: a) $\mu^{[j](t+1)} = \frac{\sum_{i=1}^n q(z^i)x^i}{\sum_{i=1}^n q(z^i)}$ b) $\pi^{[j](t+1)} = \frac1n\sum_{i=1}^n q(z^i)$ 4) Repeat

Further proofs *M* — Likelihood: $L = \sum_{i=1}^n\log(\sum_{j=1}^k\pi^{[j]}N(x^i;\mu^{[j]},\Sigma^{[j]}))$ Proof of opt. mean in M-step: • $\nabla_{\mu^{[j]}}L = \sum_{i=1}^n\frac{\delta L}{\delta N(\ldots)}\times\frac{\delta N(\ldots)}{\delta\mu^{[j]}} = \sum_{i=1}^n\frac{\pi^{[j]}}{\sum_{j=1}^k\pi^{[j]}N(\ldots)}\times\frac{\delta N(\ldots)}{\delta\mu^{[j]}} = \sum_{i=1}^n\frac{\pi^{[j]}N(\ldots)}{\sum_{j=1}^k\pi^{[j]}N(\ldots)}\times\frac{\delta\log N(\ldots)}{\delta\mu^{[j]}}$ because $\frac{\delta\log N(\ldots)}{\delta\mu^{[j]}} = \frac{1}{N(\ldots)}\times\frac{\delta N(\ldots)}{\delta\mu^{[j]}}$ $= \sum_{i=1}^n\gamma^{i[j]}\times(\delta\log(\text{const.}\times\frac{1}{|\Sigma|^{1/2}}\times\exp(-\frac12(x^i-\mu^{[j]})^\top\Sigma^{-1}(x^i-\mu^{[j]}))))/(\delta\mu^{[j]}) = \sum_{i=1}^n\gamma^{i[j]}\times\frac{\delta}{\delta\mu^{[j]}}(\log(1) - \frac12\log(|\Sigma|) - \frac12(x^i-\mu^{[j]})^\top\Sigma^{-1}(x^i-\mu^{[j]})) = \sum_{i=1}^n\gamma^{i[j]}\times(-\frac12\times\frac{\delta}{\delta\mu^{[j]}}[\log(|\Sigma|) + (x^i-\mu^{[j]})^\top\Sigma^{-1}(x^i-\mu^{[j]})])$ • Then: $\sum_{i=1}^n\gamma^{i[j]}\times(-\frac12\times 2\Sigma^{-1}(x^i-\mu^{[j]})\times -1)$ due to matrix calculus $= \sum_{i=1}^n\gamma^{i[j]}\Sigma^{-1}(x^i-\mu^{[j]}) = 0$ • $\sum_{i=1}^n\gamma^{i[j]}\Sigma^{-1}x^i = \sum_{i=1}^n\gamma^{i[j]}\Sigma^{-1}\mu^{[j]}$ • $\mu^{[j]*} = \frac{\sum_{i=1}^n\gamma^{i[j]}x^i}{\sum_{i=1}^n\gamma^{i[j]}}$ Proof of opt. π in M-step: • Constr.: $\sum_{j=1}^k\pi^{[j]} = 1$ • Lag.: $\mathcal{L}(\pi^{[j]}) = \sum_{i=1}^n\log(\sum_{j=1}^k\pi^{[j]}N(\ldots)) +$

$\lambda(\sum_{j=1}^k \pi^{[j]}-1)$ • Der.:
$\nabla_{\pi^{[j]}}\mathcal{L}(\pi^{[j]})=$
$\sum_{i=1}^n \frac{\sum_{j=1}^k \mathcal{N}(\dots)}{\sum_{j=1}^k \pi^{[j]}\mathcal{N}(\dots)}+\lambda$
$=\sum_{i=1}^n \sum_{j=1}^k \frac{\gamma^{i[j]}}{\pi^{[j]}}+\lambda=0$
• $\sum_{i=1}^n\sum_{j=1}^k \gamma^{i[j]}=-\lambda\sum_{j=1}^k \pi^{[j]}$ • Sol. for $\pi^{[j]}$: $\pi^{[j]}=\frac{\sum_{i=1}^n \gamma^{i[j]}}{-\lambda}$ • Sol. for λ: $\frac{\sum_{i=1}^n\sum_{j=1}^k \pi^{[j]}\mathcal{N}(\dots)}{\sum_{i=1}^n\sum_{j=1}^k \pi^{[j]}\mathcal{N}(\dots)}=-\lambda\times 1$ due to constr. $n=-\lambda$ • Plugging λ into $\pi^{[j]}$: $\pi^{[j]*}=\frac{1}{n}\sum_{i=1}^n\gamma^{i[j]}$
L, M, E — $L\geq M\Leftrightarrow E\geq 0$:
• $E=\mathbb{E}[log(\frac{q(z^i)}{p(z^i|x^i)})]=$
$\mathbb{E}[-log(\frac{p(z^i|x^i)}{q(z^i)})]$ • Acc. to Jensen's ineq.: $E\geq -log(\mathbb{E}[\frac{p(z^i|x^i)}{q(z^i)}])=$
$-log(\sum_{i=1}^k q(z^i)\frac{p(z^i|x^i)}{q(z^i)})=$
$-log(\sum_{i=1}^k p(z^i|x^i))=-log(1)=0$
$L=M\Leftrightarrow E=0$, i.e. when $q(z^i)=p(z^i|x^i)$
Proof that EM algo conv. — • Acc. to Jensen's ineq.:
$L_t\geq\sum_{i=1}^n\sum_{z^i} q_t(z^i)\log\frac{p_{\theta_t}(x^i,z^i)}{q_t(z^i)}$
• In prev. M-step, L was max.:
$\theta_t=\arg\max_\theta(q_t(z^i)\log\frac{p(x^i,z^i)}{q_t(z^i)})$
• Thus,
$\sum_{i=1}^n\sum_{z^i} q_t(z^i)\log\frac{p_{\theta_t}(x^i,z^i)}{q_t(z^i)}\geq$
$\sum_{i=1}^n\sum_{z^i} q_t(z^i)\log\frac{p_{\theta_{t-1}}(x^i,z^i)}{q_t(z^i)}$
• RHS of this equation is output of prev. E-step, where q_t was set such that:
$\sum_{i=1}^n\sum_{z^i} q_t(z^i)\log\frac{p_{\theta_{t-1}}(x^i,z^i)}{q_t(z^i)}=$
L_{t-1} • Then $L_t\geq L_{t-1}$

Estimating Common Distributions

Gaussian *Frequentism (MLE)* —
• Likelihood: $L=(\frac{1}{\sigma})^n\prod_{i=1}^n exp(-\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2)$
• Log likelihood: $LL=-nlog(\sigma)-\sum_{i=1}^n(\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2)$
• $\mu_{MLE}=\frac{1}{n}\sum_{i=1}^n x^{(i)}$: • μ_{MLE} is an unbiased est.: ∘ $\mathbb{E}[\mu_{MLE}]=$
$\mathbb{E}[\frac{1}{n}\sum_{i=1}^n x_i]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[x_i]=\mu$
• $\sigma^2_{MLE}=\frac{1}{n}\sum_{i=1}^n(x^{(i)}-\mu)^2$:
• σ^2_{MLE} is a biased est.:
∘ $\mathbb{E}[\Sigma_{MLE}]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[(x_i-\mu_{MLE})(x_i-\mu_{MLE})^\top]=\Sigma-\frac{1}{n}\Sigma$
Bayesianism — • Assume $x\sim\mathcal{N}(\mu,\sigma^2)$ with known σ^2 and $\mu\sim\mathcal{N}(\mu_0,\sigma_0^2)$
• Posterior: $p(\mu|x,\mu_0,\sigma_0^2)\propto\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2)$ • Based on the parametric form of the Gaussian dist.:
$(n/\sigma^2+1/\sigma_0^2)\mu^2=(1/(2\sigma_P^2))\mu^2$ and
$((\sum_{i=1}^n x_i)/\sigma^2+\mu_0/\sigma_0^2)\mu=(\mu_P/\sigma_P^2)\mu$
• Solving for σ_P and μ_P:
∘ $\mu_P=(n\bar{x}\sigma_0^2+\mu_0\sigma^2)/(n\sigma_0^2+\sigma^2)$
∘ $\sigma_P=(\sigma^2\sigma_0^2)/(n\sigma_0^2+\sigma^2)$ • Conj. prior

Binomial *Frequentism* — MLE:
• Likelihood: $P(\delta|p)\sim p^s(1-p)^{n-s}$
• $\hat{p}_{\text{MLE}}=\arg\max(p^s(1-p)^{n-s})=s/n$
Bayesianism — • Post.:
$P(p|\delta)\propto p^s(1-p)^{n-s}p^{\alpha-1}(1-p)^{\beta-1}=p^{s+\alpha-1}(1-p)^{n-s+\beta-1}\sim$
Beta$(s+\alpha,n-s+\beta)$ • Conj. prior: Beta
MAP: • $\hat{p}_{\text{MAP}}=\frac{s+\alpha-1}{s+\alpha+n-s+\beta-2}$, post. mean

Poisson *Frequentism* — MLE:
• Likelihood: $P(x|\lambda)=\prod_{i=1}^n\frac{\lambda^{x_i}e^{-\lambda}}{x_i!}=$
$\frac{\lambda^{\sum_{i=1}^n x_i}e^{-\lambda n}}{\prod_{i=1}^n x_i!}$ • $\hat\lambda_{\text{MLE}}=\frac{1}{n}\sum_{i=1}^n x_i$
Bayesianism — • Post.:
$P(\lambda|x)\propto\lambda^{\sum_{i=1}^n x_i}e^{-\lambda n}\lambda^{\alpha-1}e^{-\beta\lambda}=$
$\lambda^{\sum_{i=1}^n x_i+\alpha-1}\times e^{-\lambda n}\times e^{-\beta\lambda}=$
$\lambda^{\sum_{i=1}^n x_i+\alpha-1}\times e^{-(n+\beta)\lambda}\sim$
Gamma$(\alpha+\sum_{i=1}^n x_i,\beta+n)$ • Conj. prior: Gamma MAP: • $\hat\lambda_{\text{MAP}}=\frac{\alpha+\sum_{i=1}^n x_i-1}{\beta+n}$, post. mean

Linear Regression

Form. • $y^{(i)}=\beta\cdot x^{(i)}$ resp. $y=X\beta$ • First col. of X is 1 and first el. of $\beta=\beta_0$
Opt. *Objective function* — OLSE: Minimize MSE: $LO=(y-X\beta)^\top(y-X\beta)$
Optimization —
• $\nabla_\beta LO=\frac{1}{2}\nabla_\beta((y-X\beta)^\top(y-X\beta))=$
$X^\top X\beta-X^\top y=X^\top(X\beta-y)=0$
• $\Rightarrow\beta=(X^\top X)^{-1}X^\top y$
Evaluation — • *Gauss Markov theorem*: OLSE is best (lowest variance, lowest MSE) unbia. est. Proof: ∘ Let
$\hat\beta=A^\top y=(X^\top X)^{-1}X^\top y$ be the OLSE ∘ Let $C^\top y$ be another unbia. est.
∘ $\mathbb{V}(\hat\beta)=\mathbb{V}(A^\top y)=A^\top\mathbb{V}(y)A$ since A is constant ∘ We can further develop to:
$A^\top\sigma^2 I_m A=\sigma^2A^\top A$ since var. is given by error ∘ Similarly, $\mathbb{V}(C^\top y)=\sigma^2C^\top C$
∘ For the OLSE, we can plug in $(X^\top X)^{-1}X^\top$ for A which yields:
$\mathbb{V}(A^\top y)=$
$\sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1}=$
$\sigma^2(X^\top X)^{-1}$ ∘ Then: $\mathbb{V}(A^\top y)\leq\mathbb{V}(C^\top y)$
Distr. —
• $\mathbb{E}[\beta]=(X^\top X)^{-1}X^\top(X\beta+\mathbb{E}[\epsilon])=\beta$
• $\mathbb{V}[\beta]=\mathbb{V}[(X^\top X)^{-1}X^\top y]=$
$(X^\top X)^{-1}X^\top X(X^\top X)^{-1}(\mathbb{V}[X\beta]+\mathbb{V}[\epsilon])=(X^\top X)^{-1}\sigma^2$ • $\mathbb{E}[y]=\beta^\top x+\epsilon$
• $\mathbb{V}[y]=\mathbb{V}[\epsilon]=\sigma^2$

Bayesian Linear Regression

Form. • $y^{(i)}=\beta\cdot x^{(i)}+\epsilon$ • $\beta\sim\mathcal{N}(0,T^2I_m)$ with $p(\beta)\propto-\frac{1}{2T^2}\beta^\top\beta$ • $\epsilon\sim\mathcal{N}(0,\sigma^2)$
Opt. • $p(\beta|X,y)\propto p(y|X,\beta)\times p(\beta)$:
∘ $p(\beta|X,y)\propto exp(-\frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta))\times exp(-\frac{1}{2T^2}\beta^\top\beta)=$
$exp(-\frac{1}{2}\frac{1}{\sigma^2}(y^\top y-2\beta^\top X^\top y+\beta^\top X^\top X\beta)+\frac{1}{2T^2}\beta^\top\beta)\propto$
$exp(-\frac{1}{2}(\beta^\top(\frac{1}{\sigma^2}X^\top X+\frac{1}{2T^2}I_m)\beta-\frac{2}{\sigma^2}\beta^\top X^\top y))$ We now apply a symmetric matrix property ∘ Through this, we get $p(\beta|X,y)\propto(\frac{1}{2}(\beta+(\frac{1}{\sigma^2}X^\top X+\frac{1}{T^2}I_m)^{-1}(\frac{1}{\sigma^2}X^\top y))^\top(\frac{1}{\sigma^2}X^\top X+\frac{1}{T^2}I_m)(\beta+(\frac{1}{\sigma^2}X^\top X+\frac{1}{T^2}I_m)^{-1}(\frac{1}{\sigma^2}X^\top y)))$ ∘ Thus, posterior $\sim\mathcal{N}(\mu,\Sigma)$ with ■ $\mu=\Sigma\times\frac{1}{\sigma^2}X^\top y$
■ $\Sigma=(\frac{1}{\sigma^2}X^\top X+\frac{1}{T^2}I_m)^{-1}$ ∘ Conjugate prior

Ridge (ℓ_2) Regression

Opt. *Objective function* — Lagrangian:
$LO=(y-X\beta)^\top(y-X\beta)+\lambda(\|\beta\|^2-t)$
Optimization — $\beta=(X^\top X+\lambda I)^{-1}X^\top y$
Alt. for. — Bayesian MAP:
• Posterior $p(\beta|X,y)\propto$
Likelihood $p(y|X,\beta)\sim$
$\mathcal{N}(X\beta,\sigma^2I_n)\times$ Prior $p(\beta)\sim\mathcal{N}(0,\tau^2I_m)$:
∘ $p(\beta|X,y)\propto\log(exp(-\frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta))exp(-\frac{1}{2T^2}\beta^\top\beta))=$
$-\frac{\|y-X\beta\|^2}{\sigma^2}-\frac{\|\beta\|^2}{\tau^2}$ • Max. log post. (MAP): $\arg\min_\beta\frac{\|y-X\beta\|^2}{\sigma^2}+\frac{\|\beta\|^2}{\tau^2}=$
$\|y-X\beta\|^2+\frac{\sigma^2}{\tau^2}\|\beta\|^2$ • Equiv. to log loss of regr. with $\lambda=\frac{\sigma^2}{\tau^2}$ *Effect* — Adr. multicollinearity: • SVD for $X=USV^\top$
• $X\beta=UD(D^2+\lambda I)^{-1}DU^\top Y$ Proof:
∘ $X\beta=UDV^\top((UDV^\top)^\top UDV^\top+\lambda I)^{-1}(UDV^\top)^\top Y$
∘ $=UDV^\top(VDU^\top UDV^\top+\lambda I)^{-1}VDU^\top Y$ ∘ $=$
$UDV^\top(VD^2V^\top+\lambda VV^\top)^{-1}VDU^\top Y$
∘ $=UDV^\top(V(D^2+\lambda I)V^\top)^{-1}VDU^\top Y$
∘ $=UD(D^2+\lambda I)^{-1}DU^\top Y$ • Similarly, we can show that
$\|\beta\|^2=\|V(D^2+\lambda I)^{-1}DU^TY\|^2=$
$Y^\top UD^2(D^2+\lambda I)^{-2}U^\top Y=\frac{W^\top D^2W}{(D^2+\lambda I)^2}$
where $W=U^\top Y$
Characteristics — • Strictly convex
• Global minimum • Unique sol. (lin. indep. col.) • Analy. sol. (always invert.)

Lasso (ℓ_1) Regression

Opt. *Objective function* — Lagrangian:
$LO=(y-X\beta)^\top(y-X\beta)+\lambda(|\beta|-t)$ *Alt. for.* — Bayesian MAP:
• Posterior $p(\beta|X,y)\propto$
Likelihood $p(y|X,\beta)\sim$
$\mathcal{N}(X\beta,\sigma^2I_n)\times$ Prior $p(\beta)\sim$Lapl.$(0,b)$:
∘ $p(\beta|X,y)\propto\log(exp(-\frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta))\times\prod_{j=1}^m exp(-\frac{|\beta_j|}{b}))=$
$\log p(\beta|X,y)\propto-\frac{\|y-X\beta\|^2}{\sigma^2}-\frac{\|\beta\|_1}{b}$
• Max. log post. (MAP):
$\arg\min_\beta\frac{\|y-X\beta\|^2}{\sigma^2}+\frac{\|\beta\|_1}{b}=$
$\|y-X\beta\|^2+\frac{\sigma^2}{b}\|\beta\|_1$ • Equiv. to log loss of regr. with $\lambda=\frac{\sigma^2}{b}$

Logistic Regression

Form. Sigmoid:
• $P(y=1|x)=\frac{1}{1+e^{-\beta\cdot x}}=\frac{e^{\beta\cdot x}}{1+e^{\beta\cdot x}}$ Proof:
∘ $P(y=1|x)=\frac{P(x|1)P(1)}{P(x|1)P(1)+P(x|0)P(0)}$
∘ If $p(y=0)=p(y=1)$: $=\frac{P(x|1)}{P(x|1)+P(x|0)}$
∘ Then $\beta=\Sigma^{-1}(\mu_1-\mu_0)$ and
$\beta_0=1/2(\mu_1^\top\Sigma^{-1}\mu_1-\mu_0^\top\Sigma^{-1}\mu_0)$
Opt. *Objective function* — Log likelihood:
$\log L(\beta)=$
$\sum_{i=1}^n\left[y^i\log\sigma(z^i)+(1-y^i)\log(1-\sigma(z^i))\right]$
Optimization — $\frac{\partial-\log L(\beta)}{\partial\beta}=$
$-\sum_{i=1}^n\frac{\partial}{\partial\beta}[y^i\log\sigma(z^i)+(1-y^i)\log(1-\sigma(z^i))]=\sum_{i=1}^n[\sigma(z^i)-y^i]x^i$

Bayesian Neural Networks

Form. *Variational inference*:
• Approximates true post. $p(w|D)$ by $q(w|\theta)$ • We assume $q(w|\theta)\sim\mathcal{N}(\mu,\sigma^2I)$ with $\theta=(\mu,\sigma)$
Opt. *Objective function* — • Min. KL div.:
$\theta^*=\arg\min_\theta KL[q(w|\theta)|p(w|D)]=$
$\mathbb{E}[log(\frac{q(w|\theta)}{p(w|D)})]=$
$\mathbb{E}[log(q(w|\theta))]-\mathbb{E}[log(p(w|D))]=$
$\mathbb{E}[log(q(w|\theta))]-\mathbb{E}[log(\frac{p(D|w)\times p(w)}{p(D)})]=$
$\mathbb{E}log(q(w|\theta))-\mathbb{E}log(p(D|w))-\mathbb{E}log(p(w))+\mathbb{E}log(p(D))=$
$\mathbb{E}log(q(w|\theta))-\mathbb{E}log(p(D|w))-\mathbb{E}log(p(w))$+const.=−ELBO+const. • Min. KL div. = max. ELBO
Optimization — • Let neg. ELBO
$\frac{\partial}{\partial\theta}\mathbb{E}[log(q(w|\theta))$ (entropy)$-log(p(D|w))$ (likelihood)$-log(p(w))$ (prior)$]=\frac{\partial}{\partial\theta}\mathbb{E}[F(w,\theta)]$ • Let $w=\mu+\sigma\epsilon$ • Acc. to *reparametr. trick*:
∘ $\frac{\partial}{\partial\mu}\mathbb{E}[\frac{\partial}{\partial w}F(w,\theta)+\frac{\partial}{\partial\mu}F(w,\theta)]$
∘ $\frac{\partial}{\partial\sigma}\mathbb{E}[\epsilon^\top\frac{\partial}{\partial w}F(w,\theta)+\frac{\partial}{\partial\sigma}F(w,\theta)]$

Attention • $Q=EW^q$ resp. $q_i=e_iW^q$
∘ E $(m\times h)$ ∘ W_q $(h\times d_k)$ ∘ e_i is row vector $(1\times h)$ • $K=EW^k$ ∘ E $(n\times h)$
∘ W_k $(h\times d_k)$ • $V=EW^v$ ∘ E $(n\times h)$
∘ W_v $(h\times d_v)$ • $A=\sigma(QK^\top/\sqrt{d_k})$ in $(m\times n)$ • $Z=AV$ in $(m\times d_v)$
• MHA Z=Con.$(Z_{h_1},\dots,Z_{h_h})W_O+b_O$
where ∘ Con.(…) in $(m\times(n\times n_{\text{heads}}))$
∘ W_O in $((n_{\text{heads}}\times n)\times d_v)$ ∘ b_O in $(1\times d_v)$

VAE

Form. Enc. (maps x to h), dec. (reconstr. x from h by sampl. from $q(h|x)$)
Opt. *Parameters* — θ' (prior), θ (likelihood), Φ (approx. posterior)
Objective function — • Max. mut. inf.:
$\mathbb{E}\log[p(x^i)]=$
$\sum_{i=1}^n\mathbb{E}_q\log[\frac{p(x^i,h)}{p(h|x^i)}\times\frac{q(h|x^i)}{q(h|x^i)}]=$
$\sum_{i=1}^n\mathbb{E}\log[\frac{p(x^i,h)}{q(h|x^i)}]+\mathbb{E}\log[\frac{q(h|x^i)}{p(h|x^i)}]=$
ELBO(x^i)+KL div.$(q(\cdot|x^i)|p(\cdot|x^i))$
where KL div. acts as reg. • ELBO:
∘ Lower bound of log likelihood:
$\leq\log p(x^i)$ ∘ ELBO$(x^i)=$
$\mathbb{E}\log[p(x^i|h)]+\mathbb{E}\log[\frac{p(h)}{q(h|x^i)}]=$
$\mathbb{E}\log[p(x^i|h)]-\mathbb{E}\log[\frac{q(h|x^i)}{p(h)}]$

Diff. Models

Forward — • Data: $x_0\sim p(x)$ • Noise: $x_T\sim\mathcal{N}(0,\sigma^2I)$
• $x_0\rightarrow x_1\rightarrow\dots\rightarrow x_T$ • Stochastic diff. equation: $dx=f(x,t)dt+g(t)dW$
∘ $f(x,t)$: Det. *drift term* ∘ $g(t)$: *Diff. coeff.*: Scales noise ∘ dW: Random *Brownian motion* • 2 transitions:
∘ $q(x_t|x_{t-1})\sim\mathcal{N}(x_t|\sqrt{1-\beta_t}x_{t-1},\beta_tI)$
where β_t is added noise
∘ $q(x_t|x_0)\sim\mathcal{N}(x_t|\sqrt{\bar\alpha_t}x_0,(1-\bar\alpha_t)I)$
where $\alpha_t=1-\beta_t$ is ret. signal and $\bar\alpha_t=\prod_{s\leq t}\alpha_t$ is cum. ret. signal until t
Reverse — • $x_T\rightarrow x_{T-1}\rightarrow\dots\rightarrow x_0$
• $dx=[f(x,t)-g(t)^2\nabla_x\log p(x,t)]dt+g(t)dW$ ∘ $\nabla_x\log p(x,t)$: Score ∘ $g(t)^2$: Scales score • To approx. the score resp. pred. the noise, we train
$f_\theta(x,t)\approx\nabla_x\log p(x,t)$ Proof: ∘ Gradient of log prob. of Gaussian is: $\frac{-(x-\mu)}{\sigma^2}$
∘ From this and $q(x_t|x_0)$ from forward diff., score is:
$\nabla_x\log p(x,t)=\frac{-(x_t-\sqrt{\bar\alpha_t}\hat{x}_0)}{(1-\bar\alpha_t)}$
∘ $x_t-\sqrt{\bar\alpha_t}\hat{x}_0=f_\theta(x,t)$ is the pred. noise
∘ So $f_\theta(x,t)$ is equal to $\nabla_x\log p(x,t)$ up to a factor $-\frac{1}{(1-\bar\alpha_t)}$ • $f_\theta(x,t)$ is trained to min. delta betw. added and pred. noise:
$min_\theta\mathbb{E}\|\epsilon-f_\theta(x+\sigma_t\epsilon,t)\|^2$ • Reconstr. formula:
$x_{t-1}=\frac{1}{\sqrt{\alpha_t}}(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}}\epsilon_\theta(x_t,t))+\sigma_tz$
∘ $\frac{1}{\sqrt{\alpha_t}}$: Scaling factor
∘ $\frac{1-\alpha_t}{\sqrt{1-\bar\alpha_t}}\epsilon_\theta(x_t,t)$: Pred. noise

Kernel Methods

• $k(x^i,x^j)=\varphi(x^i)\cdot\varphi(x^j)\Leftrightarrow$ kernel func. is psd ($\alpha^\top K\alpha=\|\sum_i\alpha^i\varphi(x^i)\|^2\geq0$) and symmetric (flip arg.)
Form. • Feature map $\varphi:\mathbb{R}^m\rightarrow\mathbb{R}^k$
• $\beta\cdot\varphi(x^i)$ • Cost func.:
$LO=\sum_{i=1}^n LO(y^i,\beta\cdot\varphi(x^i)+\Omega(\beta))$ • Iff $\Omega(\beta)$ is a non-decreasing func., then:
$\beta=\Phi^\top\alpha=\sum_{i=1}^n\alpha^i\varphi(x^i)$ • Pred.:
$\beta\cdot\varphi(z)=\Phi^\top\alpha\cdot\varphi(z)=$
$\sum_{i=1}^n\alpha^i\varphi(x^i)\cdot\varphi(z)=\sum_{i=1}^n\alpha^ik(x^i,z)$
∘ Pred. measures sim. to train. inst.

Kernel Types

RBF — • $k(x^i,x^j)=\sigma^2exp(-\frac{\|x^i-x^j\|^2}{2l^2})$
Periodic — • $k(x^i,x^j)=$
$\sigma^2exp(-\frac{2\sin^2(\pi|x^i-x^j|/p)}{l^2})$
$\sigma\rightarrow$higher ampl.,$l\rightarrow$smoother,$p\rightarrow$longer per.
Kernel comp. — • $k'(x_1,x_2)=ck(x_1,x_2)$, since $\varphi'(x)=\sqrt{c}\varphi(x)$
• $k'(x_1,x_2)=f(x_1)k(x_1,x_2)f(x_2)$, since $\varphi'(x)=f(x)\varphi(x)$
• $k'(x_1,x_2)=k_1(x_1,x_2)+k_2(x_1,x_2)$, since $\varphi'(x)=[\varphi_1(x)\quad\varphi_2(x)]^\top$
• $k'(x_1,x_2)=k_1(x_1,x_2)k_2(x_1,x_2)$, since $\varphi'(x)=\varphi_1(x)\otimes\varphi_2(x)$ (elw. tensor prod.)
• $k'(x_1,x_2)=k(\psi(x_1),\psi(x_2))$, since $\varphi'(x)=\varphi(\psi(x))$
• $k'(x_1,x_2)=p(k(x_1,x_2))$ where p is a polyn., since $=\sum_i a_ik_i(x_1,x_2)$ cf. above
• $k'(x_1,x_2)=\exp(k(x_1,x_2))$, since acc. to Taylor exp. $\sum_{n=1}^r\frac{k(x_1,x_2)^r}{r!}=exp(k(x_1,x_2))=k'(x_1,x_2)$ cf. above
• $k'(x_1,x_2)=x_1^\top Ax_2$ for psd, symm. $A=L^\top L$ acc. to Chol. decomp.

GP

Form. • $y^i=\beta\cdot x^i+\epsilon$ • $\beta\sim\mathcal{N}(0,\Lambda^{-1})$
• $\epsilon\sim\mathcal{N}(0,\sigma I_m)$
Opt. • Mean and var.:
• $\mathbb{E}[y]=X\mathbb{E}(\beta)=X0=0$ ∘ Cov$(y)=\mathbb{E}[(X\beta+\epsilon)(X\beta+\epsilon)^\top]-0=X\mathbb{E}(\beta\beta^\top)X^\top+X\mathbb{E}(\beta)\mathbb{E}(\epsilon^\top)+\mathbb{E}(\epsilon)\mathbb{E}(\beta^\top)X+\mathbb{E}(\epsilon\epsilon^\top)$
where ■ $\mathbb{E}(\beta\beta^\top)=\mathbb{V}(\beta)$ and $\mathbb{E}(\epsilon\epsilon^\top)=\mathbb{V}(\epsilon)$ because zero-mean var.
∘ Plugging in the variance for β and ϵ, we get Cov$(y)=X\Lambda^{-1}X^\top+\sigma^2I_m$ ∘ This can be written as a Kernel matrix K with
$K_{ij}=x^{i\top}\Lambda^{-1}x^j+\sigma^2$ • $f\sim\mathcal{GP}(\mu,K)$
• Pred.:
$p(y_{n+1})\sim\mathcal{N}(k^\top C_n^{-1}y,c-k^\top C_n^{-1}k)$
where ∘ $k=$
$[k(x^{(1)},x^{(n+1)}),\dots,k(x^{(n)},x^{(n+1)})]^\top\equiv$
$[\varphi(x^{(1)})\cdot\varphi(x^{(n+1)}),\dots,\varphi(x^{(n)})\cdot\varphi(x^{(n+1)})]^\top$ is the kernel vector
∘ $C_n=k(x^i,x^j)+\sigma^2I_m$
∘ $c=k(x^{(n+1)},x^{(n+1)})+\sigma^2I_m$ Proof:
∘ Joint dist. $p(\begin{bmatrix}y\\y_{n+1}\end{bmatrix})\sim\mathcal{N}(0,\begin{bmatrix}C_n&k\\k^\top&c\end{bmatrix})$
∘ Use theorem: ■ Given joint dist.:
$\begin{bmatrix}a_1\\a_2\end{bmatrix}\sim\mathcal{N}(\begin{bmatrix}u_1\\u_2\end{bmatrix},\begin{bmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{bmatrix})$ ■ Cond. dist.: $p(a_2|a_1=z)=$
$\mathcal{N}(u_2+\Sigma_{21}\Sigma_{11}^{-1}(z-u_1),\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$
∘ Then:
$p(y_{n+1})=\mathcal{N}(k^\top C_n^{-1}y,c-k^\top C_n^{-1}k)$
• Pred.:
$p(y_*)\sim\mathcal{N}(K_{*n}(K_{nn}+\sigma^2)^{-1}y_n,K_{**}-K_{*n}(K_{nn}+\sigma^2)^{-1}K_{n*})$ where ∘ $*$ = new inst., n = old inst. ∘ E.g. K_{*n} has new inst. on rows and old inst. on cols
• $\int p(y_*|y_n)p(y_n|x)=p(y_*|x)$. Proof:
$\int\mathcal{N}$(a;Bc,D)$\mathcal{N}$(c;e,F)dc∼$\mathcal{N}$(a;Be,FD+BFB$^\top$)
Algo — 1) Comp. kernel matrix based on obs. data 2) Comp. kernel vector based on obs. data and new inst. 3) Calc. mean and variance of pred. dist. 4) Return pred. dist.
Proofs *Combining GPs* — • $f\sim$GP(m,k) with cov. k_1 • If we place a prior on the mean $m(x)=a^\top x+b$, with $a\sim\mathcal{N}(0,\sigma_a^2)$ and $b\sim\mathcal{N}(0,\sigma_b^2)$, then
$f\sim$GP$(0,k=k_1+k_2=k_1+\sigma_a^2x^\top x'+\sigma_b^2)$:
∘ $\mathbb{E}[m(x)]=\mathbb{E}[a]=\mathbb{E}[b]=0$
∘ $k_2(m(x),m(x'))=$Cov$[m(x),m(x')]=$
$\mathbb{E}[m(x)m(x')]-\mathbb{E}[m(x)]\mathbb{E}[m(x')]=$
$\mathbb{E}[(a^\top x+b)(a^\top x'+b)]-0=$
$\mathbb{E}[(a^\top x)(a^\top x')]+\mathbb{E}[b^2]+\mathbb{E}[ba^\top(x+x')]=\sigma_a^2x^\top x'+\sigma_b^2+0$ as b,a are indep.

Kernel Ridge Regr.

Form. $y=\beta\cdot\varphi(x)$
Opt. • Primal solution: ∘ Parameters:
$\beta=(\Phi^\top\Phi+\lambda I)^{-1}\Phi^\top y$ ∘ Pred.:
$\beta\cdot\varphi(z)=(\Phi^\top\Phi+\lambda I)^{-1}\Phi^\top y\cdot\varphi(z)=$
$y^\top\Phi(\Phi^\top\Phi+\lambda I)^{-1}\varphi(z)$ ∘ Define $K=\Phi\Phi^\top$ with $K_{ij}=\varphi(x^{(i)})\cdot\varphi(x^{(j)})$
• Dual solution: ∘ Parameters:
$\beta=\Phi^\top\alpha=\Phi^\top(K+\lambda I)^{-1}y$ Proof:
■ Theorem: $(FH^{-1}G+E)^{-1}FH^{-1}=E^{-1}F(GE^{-1}F+H)^{-1}$ ■ With $E=I_D,F=\Phi^\top,G=\Phi,H=I_N$, simplify to
$(\Phi^\top\Phi+I_D)^{-1}\Phi^\top=$
$I_D^{-1}\Phi^\top(\Phi\Phi^\top+I_N)^{-1}$ ■ Since $I_D^{-1}=I_D$, we have:
$(\Phi^\top\Phi+I_D)^{-1}\Phi^\top=\Phi^\top(\Phi\Phi^\top+I_N)^{-1}$
■ $\Rightarrow(\Phi^\top\Phi+I_D)^{-1}\Phi^\top y=\beta=\Phi^\top(\Phi\Phi^\top+I_N)^{-1}y=\sum_i\varphi(x^i)[(\Phi\Phi^\top+I_N)^{-1}y]_i$
∘ Pred.: $\beta\cdot\varphi(z)=\Phi^\top\alpha\cdot\varphi(z)=$
$y^\top(\Phi\Phi^\top)^{-1}\Phi\varphi(z)=y^\top(\Phi\Phi^\top)^{-1}k$
where
$k=\Phi\varphi(z)=[k(x^{(1)},z),\dots,k(x^{(n)},z)]^\top=[\varphi(x^{(1)})\cdot\varphi(z),\dots,\varphi(x^{(n)})\cdot\varphi(z)]^\top$

Kernel SVM

Opt. • Cost func.
$\arg\min_\beta\frac{1}{2}\|\beta\|^2+C\sum_{i=1}^n\xi^i$ resp. dual Lag. obj. func. $\arg\max_\alpha\sum_{i=1}^n\alpha^i-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n\alpha^i\alpha^jy^iy^j\varphi(x^i)\cdot\varphi(x^j)$
where α = Lag. multiplier • Eligible for kernel trick • Thus, $\beta^*=\sum_{i=1}^n\tilde\alpha^i\varphi(x^i)$
where $\tilde\alpha$ = Kernel theorem • Given general Lag. solution, $\beta^*=\sum_{i=1}^n\alpha^iy^i\varphi(x^i)$, we know that: $\tilde\alpha^i=\alpha^iy^i$ • Pred.:
$\beta^*\cdot\varphi(z)=\sum_{i=1}^n\tilde\alpha^i\varphi(x^i)\cdot\varphi(z)=$
$\sum_{i=1}^n\alpha^iy^i\varphi(x^i)\cdot\varphi(z)$
Proofs SVMs are 1-nearest-neigh. clf. for RBF kernel: • Inst. x with nearest neigh. x_p and the 2nd nearest neigh. x_q • Pred.:
$f(x)=$sign$(\sum_{i=1}^n\alpha^iy^ik(x^i,x))$ • With α^i=1 RBF kernel:
$f(x)=$sign$(\sum_{i=1}^n y^i\exp(-\frac{\|x-x^i\|^2}{h^2}))$
$=$sign$(y_p\exp(-\frac{\|x-x_p\|^2}{h^2})+$
$\sum_{j=1,j\neq p}^n y^j\exp(-\frac{\|x-x_j\|^2}{h^2}))$ • If conditions on h guarantee that term 1 > term 2, then $f(x)$=sign(y_p) and SVM is 1-nearest-neigh. • Term 2:
$\sum_{j=1,j\neq p}^n y^j\exp(-\frac{\|x-x_j\|^2}{h^2})\leq$
$(n-1)\exp(-\frac{\|x-x_q\|^2}{h^2})$ since x_q is the 2nd-nearest neigh. • Term 1:
$\left|y_p\exp(-\frac{\|x-x_p\|^2}{h^2})\right|=\exp(-\frac{\|x-x_p\|^2}{h^2})$
since output label $y\in\{-1,1\}$ • Then we can set: $\exp(-\frac{\|x-x_p\|^2}{h^2})>$
$(n-1)\exp(-\frac{\|x-x_q\|^2}{h^2})$
$\Rightarrow\exp(\frac{\|x-x_q\|^2-\|x-x_p\|^2}{h^2})>(n-1)$
$\Rightarrow\frac{\|x-x_q\|^2-\|x-x_p\|^2}{\log(n-1)}>h\Rightarrow h_0>h$

Hard-Margin SVM Clf.

Form. • $y\in\{-1,1\}$ • Discriminant:
∘ f=sgn$(\beta\cdot x+b)$ ∘ If sgn$_+$, f=1, else −1
∘ Lin. sep. hyperpl. given by $z=\beta\cdot x+b=0$
∘ $z\perp\beta$ • Single margin: $\gamma^i=\frac{y^i(\beta\cdot x^i+b)}{\|\beta\|}$
• For well-classified points, $\gamma^i>0$, for mis-classified points, $\gamma^i<0$ • Margin is scaling invariate. Thus, $\gamma=\frac{1}{\|\beta\|}$ • Double margin: $\gamma=\frac{2}{\|\beta\|}$
Opt. *Parameters* — Find β and b *Objective func.* — • Max. $\gamma=\frac{2}{\|\beta\|}$ (resp. $2m$) s.t.
$y^i(\beta\cdot x^i+b)\geq1$ • Min. $\gamma=\frac{1}{2}\|\beta\|^2$ s.t.
$1-y^i(\beta\cdot x^i+b)\leq0$ • Lag.:
$\mathcal{L}=\frac{1}{2}\|\beta\|^2+\sum_{i=1}^n\alpha^i(1-y^i(\beta\cdot x^i+b))$
Opt. — • Gen. sol.:
∘ $\nabla_\beta\mathcal{L}=\beta-\sum_{i=1}^n\alpha^iy^ix^i=0$
∘ $\Rightarrow\beta^*=\sum_{i=1}^n\alpha^iy^ix^i$
∘ $\nabla_b\mathcal{L}=-\sum_{i=1}^n\alpha^iy^i=0$ ∘ S.t.:
■ $1-y^i(\beta\cdot x^i+b)\leq0$ ■ $\alpha^i\geq0$
■ $\alpha^i(1-y^i(\beta\cdot x^i+b))=0$ • By Slater's cond. $(\exists(\beta,b)$ s.t. $y^i(\beta\cdot x^i+b)>1)$, strong duality holds • Obj. func. in dual Lag., after plugging in β:
$\mathcal{D}=\frac{1}{2}(\sum_{i=1}^n\sum_{j=1}^n\alpha^i\alpha^jy^iy^jx^i\cdot x^j)+$
$\sum_{i=1}^n\sum_{j=1}^n\alpha^i-\alpha^i\alpha^jy^iy^jx^i\cdot x^j=$
$\sum_{i=1}^n\alpha^i-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n\alpha^i\alpha^jy^iy^jx^i\cdot x^j$ • Dual Opt.: Maximize α s.t. ∘ $\alpha^i\geq0$
∘ $\sum_{i=1}^n\alpha^iy^i=0$ due to $\nabla_b\mathcal{L}$ • Only *support vectors* ($\alpha^i>0$, on hyperpl. $y^i(\beta\cdot x^i+b)=1$) matter for β^* and b^*:
∘ Based on compl. slack. cond.
$\alpha^i(1-y^i(\beta\cdot x^i+b))=0$: Either ■ $\alpha^i=0$ and $(1-y^i(\beta\cdot x^i+b))>0$ resp. $y^i(\beta\cdot x^i+b)>1$ or ■ $\alpha^i>0$ and $(1-y^i(\beta\cdot x^i+b))=0$ resp. $y^i(\beta\cdot x^i+b)=1$

Soft-Margin SVM Clf.

Opt. *Objective func.* — • Hinge loss:
$max(0,1-y^i)$ • Corr. ineq. constr. with slack var.: $y^i(\beta\cdot x^i+b)\geq1-\xi^i$ • For
■ Well-classified inst. outs. mar. $\xi^i<0$
■ Well-classified inst. in mar. $0<\xi^i<1$
■ Mis-classified inst. $\xi^i>1$ • Cost func. with slack var. penalized by ℓ_1 norm:
$\frac{1}{2}\|\beta\|^2+C\sum_{i=1}^n\xi^i$ where $\|\beta\|^2$ max. margin and $C\sum_{i=1}^n\xi^i$ min. hinge loss (→ hard-margin if $C\rightarrow\infty$) s.t.:
∘ $1-\xi^i-y^i(\beta\cdot x^i+b)\leq0$ ∘ $-\xi^i\leq0$ • Lag.:
$\mathcal{L}=\frac{1}{2}\|\beta\|^2+C\sum_{i=1}^n\xi^i+\sum_{i=1}^n\alpha^i(1-\xi^i-y^i(\beta\cdot x^i+b))-\sum_{i=1}^n\zeta^i\xi^i$ *Opt.* — • Gen. sol.: ∘ $\nabla_\beta\mathcal{L}=\beta-\sum_{i=1}^n\alpha^iy^ix^i=0$
∘ $\Rightarrow\beta^*=\sum_{i=1}^n\alpha^iy^ix^i$
∘ $\nabla_b\mathcal{L}=-\sum_{i=1}^n\alpha^iy^i=0$
∘ $\nabla_{\xi^i}\mathcal{L}=C-\alpha^i-\zeta^i=0$ ∘ S.t.:
■ $1-\xi^i-y^i(\beta\cdot x^i+b)\leq0$ ■ $-\xi^i\leq0$
■ $\alpha^i,\zeta^i\geq0$ ■ $\alpha^i(-\xi^i-y^i(\beta\cdot x^i+b)+1)=0$
■ $\zeta^i(-\xi^i)=0$ ∘ ... plug in found β,b,ξ^i ... • Max. α s.t. $\sum_{i=1}^n\alpha^{(i)}y^{(i)}=0$ due to $\nabla_b\mathcal{L}$ ∘ $0\leq\alpha^{(i)}\leq C$ due to $\nabla_{\xi^{(i)}}\mathcal{L}$

Extensions to the SVM

Multiclass SVMs *Form.* — • z is class of x^i • $(\beta_{z^i}\cdot x^i+b_{z^i})-(\beta_z\cdot x^i+b_z)\geq1$ for $\forall x^i$ and $\forall z\neq z^i$
Opt. — • Min.: $\frac{1}{2}\|B\|^2+C\sum_{i=1}^n\xi^i=\frac{1}{2}\sum_{z=1}^k\|\beta_z\|^2+C\sum_{i=1}^n\xi^i$ s.t.:
∘ $1-\xi^i-(\beta_{z^i}\cdot x^i+b_{z^i})+(\beta_z\cdot x^i+b_z)\leq0$ for $\forall z\neq z^i$ ∘ $-\xi^i\leq0$
Structured SVMs *Form.* — • Feature func. $\Psi(y,x)$ • Scoring func. $f(y,x)=\beta\cdot\Psi(y,x)$ • Clf. via $\hat{y}=\arg\max_y f(y,x)$ • Then:
$f(y,x)-\max_{y'}f(y',x)=$
$\beta\cdot\Psi(y,x)-\max_{y'}\beta\cdot\Psi(y',x)\geq1$ for $\forall x^i$
Opt. — • Cost func.: $\frac{1}{2}\|B\|^2+C\sum_{i=1}^n\xi^i=$
$\frac{1}{2}\sum_{z=1}^k\|\beta_z\|^2+C\sum_{i=1}^n\xi^i$ s.t.:
∘ $-\xi^i-\beta\cdot\Psi(y,x)+\max_{y'}[\Delta(y,y')+\beta\cdot\Psi(y',x)]\leq0$ resp.
$\Delta(y,y')-\xi^i-\beta\cdot\Psi(y,x)+\beta\cdot\Psi(y',x)\leq0$ for $\forall y'$ where Δ is a loss func. ∘ $-\xi^i\leq0$
• Lag.: $\mathcal{L}=\frac{1}{2}\|\beta\|^2+C\sum_{i=1}^n\xi^i+\sum_{i=1}^n\sum_y\alpha^{(ij)}(\Delta(y^i,y^j)-\xi^i-\beta\cdot\Psi(y^i,x^i)+\beta\cdot\Psi(y^j,x^i))-\sum_{i=1}^n\zeta^i\xi^i$
• We abbreviate:
$\Psi_i(y^j):=(-\Psi(y^i,x^i)+\Psi(y^j,x^i))$ and $\Delta_i(y^j)=\Delta(y^i,y^j)$ • Gen. sol.:
∘ $\nabla_\beta\mathcal{L}=\beta-\sum_{i=1}^n\sum_y\alpha^{(ij)}\Psi_i(y^j)=0$
∘ $\Rightarrow\beta^*=\sum_{i=1}^n\sum_y\alpha^{(ij)}\Psi_i(y^j)$
∘ $\nabla_b\mathcal{L}=-\sum_{i=1}^n\alpha^iy^i=0$ ∘ $\nabla_{\xi^i}\mathcal{L}=$
$\sum_{i=1}^nC-\sum_{i=1}^n\sum_y\alpha^{(ij)}-\sum_{i=1}^n\zeta^i=0$
• By Slater's cond., strong duality holds
• Obj. func. in dual Lag., after plugging in β: $\mathcal{D}=-\frac{1}{2}\|\sum_{i=1}^n\sum_y\alpha^{(ij)}\Psi_i(y^j)\|^2+\sum_{i=1}^n\sum_y\alpha^{(ij)}\Delta_i(y^j)$
Proof: ∘ $\mathcal{D}=\frac{1}{2}\|\beta^*\|^2+C\sum_{i=1}^n\xi^i+\sum_{i=1}^n\sum_y\alpha^{(ij)}\Delta(y^i,y^j)-\alpha^{(ij)}\xi^i-\alpha^{(ij)}\beta\cdot\Psi_i(y^j)-\sum_{i=1}^n\zeta^i\xi^i$
∘ $=\frac{1}{2}\|\beta^*\|^2+\sum_{i=1}^n\xi^i(C-\sum_y\alpha^{(ij)}-\zeta^i)+\sum_{i=1}^n\sum_y\alpha^{(ij)}\Delta(y^i,y^j)-\alpha^{(ij)}\beta\cdot\Psi_i(y^j)$ ∘ $(C-\sum_y\alpha^{(ij)}-\zeta^i)=0$ due to $\nabla_{\xi^i}\mathcal{L}$ ∘ Plug in β and contr. first 2 terms:
$=-\frac{1}{2}\|\beta^*\|^2+\sum_{i=1}^n\sum_y\alpha^{(ij)}\Delta(y^i,y^j)=$
$-\frac{1}{2}\|\sum_{i=1}^n\sum_y\alpha^{(ij)}\Psi_i(y^j)\|^2+\sum_{i=1}^n\sum_y\alpha^{(ij)}\Delta_i(y^j)$ • Dual Opt.:
Maximize α s.t. $0\leq\sum_y\alpha^{(ij)}\leq C$ due to $\nabla_{\xi^i}\mathcal{L}$
SVMs in Practice — For a structured SVM, we need to define 4 functions: • Feature func. $\Psi(y,x)$ • Loss func. $\Delta(y',y)$
• Prediction rule $\arg\max_y\beta\cdot\Psi(y,x)$
• Loss-augmented inf.
$\arg\max_{y'}(\Delta(y',y^i)+\beta\cdot\Psi(y',x^i))$
SVM Regressor *Form.* — • ϵ is margin
• $\hat\xi^i$ and ξ^i are slack var. • Then:
$(\beta\cdot x^i+b)-y^i\leq\epsilon+\hat\xi^i$ and
$y^i-(\beta\cdot x^i+b)\leq\epsilon+\xi^i$ for all x^i
Opt. — • Min.: $\frac{1}{2}\|B\|^2+C\sum_{i=1}^n(\xi^i+\hat\xi^i)$
s.t.: ∘ $(\beta\cdot x^i+b)-y^i-\epsilon-\hat\xi^i\leq0$
∘ $y^i-(\beta\cdot x^i+b)-\epsilon-\xi^i\leq0$
∘ $-\xi^i\leq0$ ∘ $-\hat\xi^i\leq0$ • Lag.: $\mathcal{L}=$
$\frac{1}{2}\|B\|^2+C\sum_{i=1}^n(\xi^i+\hat\xi^i)-\sum_{i=1}^n\hat\mu^i\hat\xi^i-\sum_{i=1}^n\mu^i\xi^i+\sum_{i=1}^n\hat\alpha^i((\beta\cdot x^i+b)-y^i-\epsilon-\hat\xi^i)+\sum_{i=1}^n\alpha^i(y^i-(\beta\cdot x^i+b)-\epsilon-\xi^i)$
• Gen. sol.:
∘ $\nabla_\beta\mathcal{L}=\beta+\sum_{i=1}^n\hat\alpha^ix^i-\sum_{i=1}^n\alpha^ix^i=\beta+\sum_{i=1}^n(\hat\alpha^i-\alpha^i)x^i=0$
∘ $\Rightarrow\beta^*=\sum_{i=1}^n\sum_{i=1}^n(\alpha^i-\hat\alpha^i)x^i$ ∘ $\nabla_b\mathcal{L}=\sum_{i=1}^n\hat\alpha^i-\sum_{i=1}^n\alpha^i=\sum_{i=1}^n(\hat\alpha^i-\alpha^i)=0$
∘ $\nabla_{\xi^i}\mathcal{L}=\sum_{i=1}^nC-\sum_{i=1}^n\mu^i-\sum_{i=1}^n\alpha^i=0$,
analog for $\hat\xi^i$ • By Slater's cond., strong duality holds • Obj. func. in dual Lag., after plugging in β:
$\mathcal{D}=-\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(\alpha^i-\hat\alpha^i)(\alpha^j-\hat\alpha^j)x^{(j)\top}x^i+\sum_{i=1}^n(\alpha^i-\hat\alpha^i)y^i-\epsilon\sum_{i=1}^n(\alpha^i+\hat\alpha^i)$
Proof: ∘ 1) $\mathcal{D}=C\sum_{i=1}^N(\xi^i+\hat\xi^i)$ 2)
$+\frac{1}{2}\sum_{i=1}^N\left((\alpha^i-\hat\alpha^i)x^i\right)^T\left(\sum_{i=1}^N(\alpha^i-\hat\alpha^i)x^i\right)$ 3) $-\sum_{i=1}^N\mu^i\xi^i-\hat\mu^i\hat\xi^i$ 4)
$+\sum_{i=1}^N(\alpha^i-\hat\alpha^i)y^i$ 5) $+\sum_{i=1}^N(\hat\alpha^i-\alpha^i)b$ 6) $+\sum_{i=1}^N(-\hat\alpha^i-\alpha^i)\epsilon$ 7)
$-\sum_{i=1}^N\hat\alpha^i\hat\xi^i-\alpha^i\xi^i$ 8)
$-\sum_{i=1}^N\alpha^i\sum_{j=1}^N((\alpha^j-\hat\alpha^j)x^j)^Tx^i$
$+\sum_{i=1}^N\hat\alpha^i\sum_{j=1}^N((\alpha^j-\hat\alpha^j)x^j)^Tx^i$ • We can simplify: 1, 3, 7) $\mathcal{D}=\sum_{i=1}^N\left[C-\mu^i-\alpha^i\right]\xi^i+\sum_{i=1}^N\left[C-\hat\mu^i-\hat\alpha^i\right]\hat\xi^i$ 8)
$-\sum_{i=1}^N\sum_{j=1}^N(\alpha^i-\hat\alpha^i)(\alpha^j-\hat\alpha^j)x^{(j)\top}x^i$
2) $+\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N(\alpha^i-\hat\alpha^i)(\alpha^j-\hat\alpha^j)x^{(j)\top}x^i$ 4) $+\sum_{i=1}^N(\alpha^i-\hat\alpha^i)y^i$ 5)
$+\sum_{i=1}^N(\hat\alpha^i-\alpha^i)b$ 6) $+\sum_{i=1}^N(-\hat\alpha^i-\alpha^i)\epsilon$
∘ Derivatives and certain terms $((\sum_{i=1}^n\alpha^i-\hat\alpha^i),(C-\mu^i-\alpha^i),(C-\hat\mu^i-\hat\alpha^i))$ in 1,3,5,7,8) simplify to 0