$\mathbb{E}[AX+b]=A\mathbb{E}[X]+b, \mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y]$, if $X\perp Y$: $\mathbb{E}[XY^T]=\mathbb{E}[X]\mathbb{E}[Y]^T, \mathbb{E}[X|Y=y]\doteq\int xp_{X|Y}(x|y)dx$
$\mathbb{E}_Y[\mathbb{E}_X[X|Y]]=E[X], \mathbb{E}[g(X)]=\int g(x)\cdot p(x)dx$,
$\mathrm{Cov}[X,Y]\doteq\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])^T]=\mathbb{E}[XY^T]-\mathbb{E}[X]\mathbb{E}[Y]^T$,
$\mathrm{Cov}[X,Y]=\mathrm{Cov}[Y,X]^T, \mathrm{Cov}[AX+c,BY+d]=A\mathrm{Cov}[X,Y]B^T$
$X\perp Y\Rightarrow\mathrm{Cov}[X,Y]=0, \mathrm{Var}[X]\doteq\mathrm{Cov}[X,X], \mathrm{Var}[AX+b]=A\mathrm{Var}[X]A^T, \mathrm{Var}[X+Y]=\mathrm{Var}[X]+\mathrm{Var}[Y]-2\mathrm{Cov}[X,Y]$,
$\mathrm{Var}[X|Y]\doteq\mathbb{E}[(X-\mathbb{E}[X|Y])(X-\mathbb{E}[X|Y])^T\mid Y]$
$\mathrm{Var}[X]=\mathbb{E}_Y[\mathrm{Var}_X[X|Y]]+\mathrm{Var}_Y[\mathbb{E}_X[X|Y]]$
C.of variable: X with density p_X and g diff.able and invertible function: then $Y=g(X), p_Y(y)=p_X\big(g^{-1}(y)\big)\cdot\big|\det(Dg^{-1}(y))\big|$
$\mathcal{N}(x;\mu,\sigma^2)\doteq\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-((x-\mu)^2)/(2\sigma^2))$
$\mathcal{N}(x;\mu,\Sigma)\doteq1/\Big(\sqrt{(2\pi)^n\det(\Sigma)}\Big)\exp\big(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\big)$
$\log\mathcal{N}(x;\mu,\Sigma)=-(1/2)x^t\Sigma^{-1}x+\mu^T\Sigma^{-1}x+\text{const}$
Gaussians are independent $\Leftrightarrow$ their Cov $=0$. For *index sets* A, B: $X_A\sim\mathcal{N}(\mu_A,\Sigma_{AA})$, and $X_A\mid X_B=x_b\sim\mathcal{N}\big(\mu_{A|B},\Sigma_{A|B}\big)$ where
$\mu_{A|B}\doteq\mu_A+\Sigma_{AB}\Sigma_{BB}^{-1}(x_B-\mu_B), \Sigma_{A|B}\doteq\Sigma_{AA}-\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$
$AX+b\sim\mathcal{N}(A\mu+b,A\Sigma A^T), X=\Sigma^{1/2}Y+\mu$ for $Y\sim\mathcal{N}(0,I)$
X and X' **indpendent** GRVs then $X+X'\sim\mathcal{N}(\mu+\mu',\Sigma+\Sigma')$
For jointly Gaussian RVs X_A, X_B, independent $\varepsilon\sim\mathcal{N}(0,\Sigma_{A|B})$:
$X_A\doteq AX_B+b+\varepsilon$ for $A\doteq\Sigma_{AB}\Sigma_{BB}^{-1}$, and $b\doteq\mu_A-\Sigma_{AB}\Sigma_{BB}\mu_B$
(*Jensen's*): convex function $g:\mathbb{R}\to\mathbb{R}$, (e.g. S[u]) $g(\mathbb{E}[X])\le\mathbb{E}[g(X)]$
$S[u]\doteq-\log(u), H[p]\doteq\mathbb{E}_{x\sim p}[S[p(x)]]=\mathbb{E}_{x\sim p}[-\log p(x)]$
$H[p\|q]\doteq\mathbb{E}_{x\sim p}[S[q(x)]]=\mathbb{E}_{x\sim p}[-\log q(x)]=H[p]+\mathrm{KL}(p\|q)$
$\mathrm{KL}(p\|q)\doteq H[p\|q]-H[p]=\mathbb{E}_{\theta\sim p}[S[q(\theta)]-S[p(\theta)]]=\mathbb{E}_{\theta\sim p}\big[\log\frac{p(\theta)}{q(\theta)}\big]$
(*Gibb's*): $\mathrm{KL}(p\|q)\ge0$, $\mathrm{KL}(p\|q)=0\Leftrightarrow p=q$ a.s., $\mathrm{KL}(p\|q)\ne\mathrm{KL}(q\|p)$
$H[X\mid Y=y]\doteq\mathbb{E}_{x\sim p(x|y)}[-\log p(x|y)]$,
$H[X\mid Y]\doteq\mathbb{E}_{y\sim p(y)}[H[X\mid Y=y]]=\mathbb{E}_{(x,y)\sim p(x,y)}[-\log p(x|y)]$
$H[X,Y]\doteq\mathbb{E}_{(x,y)\sim p(x,y)}[-\log p(x,y)]$
$H[X,Y]=H[Y]+H[X\mid Y]=H[X]+H[Y\mid X]=H[Y,X]$
Bayes' rule for entropy: $H[X\mid Y]=H[Y\mid X]+H[X]-H[Y]$
"Information never hurts" principle: $H[X\mid Y]\le H[X]$
$0\le I(X;Y)\doteq H[X]-H[X\mid Y]=H[X]+[Y]-H[X,Y]=I(Y;X)$
$I(X;Y|Z)\doteq H[X|Z]-H[X|Y,Z]=I(Y;X|Z)=I(X;Y,Z)-I(X;Z)$

0.1 Information theoretic quantities about Gaussians

Entropy $H[\mathcal{N}(\mu,\Sigma)]=\frac{1}{2}\log\det(2\pi e\Sigma)=\frac{1}{2}\log\big((2\pi e)^d\det(\Sigma)\big)$
KL $p\doteq\mathcal{N}(\mu_p,\Sigma_p)$ and $q\doteq\mathcal{N}(\mu_q,\Sigma_q)$, then $\mathrm{KL}(p\|q)=$
$\frac{1}{2}\Big(\mathrm{tr}\big(\Sigma_q^{-1}\Sigma_p\big)+\big(\mu_p-\mu_q\big)^T\Sigma_q^{-1}\big(\mu_p-\mu_q\big)-d+\log\frac{\det(\Sigma_q)}{\det(\Sigma_p)}\Big)$
M.I. Let $X\sim\mathcal{N}(\mu,\Sigma)$ and $Y=X+\varepsilon$ where $\varepsilon\sim\mathcal{N}(0,\sigma_n^2I)$ then:
$I(X;Y)=I(Y;X)=H[Y]-H[\varepsilon]=1/2\log(\det(\sigma_n^{-2}\Sigma+I))$

1 Linear regression point estimates

$y_i=w^{*T}x_i+\varepsilon_i$ where $\varepsilon_i\sim\mathcal{N}(0,\sigma_n^2)\Leftrightarrow y_i\mid x_i,w\sim\mathcal{N}(w^Tx_i,\sigma_n^2)$
$\hat{w}_{\text{ls}}\doteq\arg\min_{w\in\mathbb{R}^d}\sum_{i=1}^n\big(y_i-w^Tx_i\big)^2=\arg\min_{w\in\mathbb{R}^d}\|y-Xw\|_2^2$
$\hat{w}_{\text{MLE}}=\arg\min_{\theta\in\Theta}-\sum_{i=1}^n\log p(y_i|x_i,\theta)=\hat{w}_{\text{LS}}=(X^TX)^{-1}X^Ty$
$\hat{w}_{\text{ridge}}\doteq\arg\min_{w\in\mathbb{R}^d}\|y-Xw\|_2^2+\lambda\|w\|_2^2=(X^TX+\lambda I)^{-1}X^Ty$
$\hat{w}_{\text{MAP}}=\arg\max_{w\in\mathbb{R}^d}\log p(y_{1:n}\mid x_{1:n},w)+\log p(w)$
$=\arg\min_{w\in\mathbb{R}^d}\|y-Xw\|_2^2+\frac{\sigma_n^2}{\sigma_p^2}\|w\|_2^2=\hat{w}_{\text{ridge}}$
Given $\hat{w}$, $f^*\doteq\hat{w}^Tx^*$ and $y^*\doteq f^*+\varepsilon, y^*\sim\mathcal{N}(f^*,\sigma_n^2I)$.

2 Bayesian linear regression (probabilistic inference)

If $w\sim\mathcal{N}(\mu_p,\Sigma_p)$, $w\mid x_{1:n},y_{1:n}\sim\mathcal{N}(\mu,\Sigma)$ where
$\Sigma\doteq\big(\sigma_n^{-2}X^TX+\Sigma_p^{-1}\big)^{-1}, \mu\doteq\Sigma\big(\Sigma_p^{-1}\mu_p+\sigma_n^{-2}X^Ty\big)=\hat{w}_{\text{MAP}}$
by the properties Gaussians, closed form predictive posterior:
$w^Tx^*\doteq f^*\sim\mathcal{N}(\mu^Tx^*,x^{*T}\Sigma x^*), y^*\sim\mathcal{N}(\mu^Tx^*,x^{*T}\Sigma x^*+\sigma_n^2I)$

3 Optimal action for given cost function C(y,a): a^*

$(y-a)^2: \mathbb{E}_y[Y|x]=\int y\cdot p(y|x)dy=\mu_{Y|X=x}$
$c_1\max(y-a,0)+c_2\max(a-y,0)$: $\mu_{Y|X=x}+\sigma_{Y|X}\Phi^{-1}(c_1/(c_1+c_2))$

4 Function-space view and non-linear regression

Given design matrix $\Phi\in\mathbb{R}^{n\times e}$ from weights prior we get:
$f\mid X\sim\mathcal{N}(\Phi\mathbb{E}[w],\Phi\mathrm{Var}[w]\Phi^T)=\mathcal{N}(0,\sigma_p^2\Phi\Phi^T)$ where
$\sigma_p^2\Phi\Phi^T\doteq K\in R^{n\times n}$ is the *kernel matrix* ($n\times n$ instead of $e\times e$).

5 Gaussian Processes

$f\sim\mathcal{GP}(\mu,k)$ for any set $A\doteq\{x_1,...,x_m\}\subseteq\mathcal{X}$, $f_A\sim\mathcal{N}(\mu_A,K_{AA})$.
$f^*|x^*\sim\mathcal{N}(\mu(x^*),k(x^*,x^*)), y^*|x^*\sim\mathcal{N}(\mu(x^*),k(x^*,x^*)+\sigma_n^2)$
For $f\sim\mathcal{GP}(\mu,k)$ can define $g(x)=f(x)-\mu(x)$, $g\sim\mathcal{GP}(0,k)$.
Can write joint **prior** distribution of $y_{1:n}$ and f^* at a test point x^* as $[y,f^*]^T\mid x^*,x_{1:n}\sim\mathcal{N}(\tilde{\mu},\tilde{K})$, where
$\tilde{\mu}\doteq\begin{bmatrix}\mu_A\\ \mu(x^*)\end{bmatrix}\tilde{K}\doteq\begin{bmatrix}K_{AA}+\sigma_n^2I & k_{x^*,A}\\ k_{x^*,A}^T & k(x^*,x^*)\end{bmatrix}k_{x,A}^T\doteq[k(x,x_1),...,k(x,x_n)]$
Conditioning on $y_{1:n}$ yields: $f\mid x_{1:n},y_{1:n}\sim\mathcal{GP}(\mu',k')$ where
$\mu'(x)\doteq\mu(x)+k_{x,A}^T(K_{AA}+\sigma_n^2I)^{-1}(y_A-\mu_A)$
$k'(x,x')\doteq k(x,x')-k_{x,A}^T(K_{AA}+\sigma_n^2I)^{-1}k_{x',A}$
Adding noise: $y^*\mid x^*,x_{1:n},y_{1:n}\sim\mathcal{N}(\mu'(x^*),k'(x^*,x^*)+\sigma_n^2)$

6 Kernel functions

- *Symmetric*: $k(x,x')=k(x',x)$ for any $x,x'\in\mathcal{X}$
- *Positive semi-definite*: K_{AA} is positive semi-def for any $A\subseteq\mathcal{X}$

Fact: a matrix is *positive (semi-)definite* $\Leftrightarrow\lambda_i>(\ge)\,0$ and $\det(A)=\prod_i\lambda_i\Rightarrow$ a matrix is *positive (semi-)definite* only if $\det(A)>(\ge)\,0$.
$k(x,x')\doteq k_1(x,x')+k_2(x,x')$ and $k(x,x')\doteq k_1(x,x')k_2(x,x')$
$k(x,x')\doteq c\cdot k_1(x,x')$ for any $c>0$ "output scale"
$k(x,x')\doteq f(k_1(x,x'))$ for $f=\exp$ or any pos. coeff. polynomial
• *stationary*: $k(x-x')=k(x,x')$ • *isotropic*: $k(\|x-x'\|)=k(x,x')$
Isotropy $\Rightarrow$ stationarity, in $\mathbb{R}\Leftrightarrow$ due to symmetry of kernel.

6.1 Maximum marginal likelihood in GP regression

$\hat{\theta}=\arg\max_\theta-\frac{n}{2}\log2\pi-\frac{1}{2}\log\det\big(K_y^{(\theta)}\big)-\frac{1}{2}y^T\big(K_y^{(\theta)}\big)^{-1}y$
$\frac{\partial}{\partial\theta_j}\log p(y_{1:n}\mid x_{1:n},\theta)=\frac{1}{2}\mathrm{tr}\Big(\big(\alpha\alpha^T-K_y^{(\theta)}\big)\frac{\partial K_y^{(\theta)}}{\partial\theta_j}\Big)\big[\alpha\doteq K_y^{(\theta)}y\big]$
In general non-convex, might have multiple local optima.

7 Laplace Approximation

$p(\theta\mid y)\approx q(\theta)=\mathcal{N}\big(\theta;\hat{\theta},\Lambda^{-1}\big)$ where $\hat{\theta}=\arg\max p(\theta\mid y)$ and $\Lambda\doteq-H_{\psi(\hat{\theta})}=-H_\theta\log p(\theta\mid x_{1:n},y_{1:n})|_{\theta=\hat{\theta}}$ is obtained by second-order Taylor of $\psi(\theta)\doteq\log p(\theta\mid x_{1:n},y_{1:n})$ around $\hat{\theta}$.

7.0.1 Laplace for Bayesian Logistic Regression

$\sigma(z)\doteq1/(1+\exp(-z))\in(0,1)$, where $z=w^Tx$, i.e.
$p(y\mid x,w)=\begin{cases}\sigma(w^Tx) & \text{if } y=1\\ 1-\sigma(w^Tx) & \text{if } y=-1\end{cases}=\sigma(y\cdot w^Tx)$
Bayesian logistic regression = BLR with $y\mid x,w\sim\mathrm{Bern}(\sigma(w^Tx))$
$\hat{w}=\arg\min_w\frac{1}{2\sigma_p^2}\|w\|_2^2+\sum_{i=1}^n\log(1+\exp(-y_iw^Tx_i)))=$
regularized logistic regression ($\log(1+\exp(-y_iw^Tx_i))=$ *logistic loss*) $\Rightarrow$ can use SGD: $\theta_{t+1}\leftarrow\theta_t-\eta_t\frac{1}{m}\sum_{i=m}^m\nabla_\theta l(\theta_t;x_i)$
$w\leftarrow w(1-2\lambda\eta_t)+\eta_tyx\sigma(-yw^Tx)$ as $\nabla_w\ell_{\log}(w^Tx;y)=-yx\sigma(-yw^Tx)$
$\Lambda=-H_w\log p(w\mid x_{1:n},y_{1:n})|_{w=\hat{w}}=\sigma_p^{-2}I+X^T\mathrm{diag}_{i\in[n]}\{\pi_i(1-\pi_i)\}X$
where $\pi_i=\sigma(\hat{w}^Tx_i)=P(Y=1\mid\hat{w},x_i)$. Making predictions:
$p(y^*\mid x^*,x_{1:n},y_{1:n})\approx\int p(y^*\mid x^*,w)q_\lambda(w)dw=\int\sigma(y^*f^*)\cdot\mathcal{N}(f^*;\hat{w}^Tx^*,x^{*T}\Lambda^{-1}x^*)df^*$ can be approximated by num. quadrature methods s.a. Gauss-Hermite quadrature.

8 Variational inference

$q^*\doteq\arg\min_{q\in\mathcal{Q}}\mathrm{KL}(q\parallel p)=\arg\min_{\lambda\in\Lambda}\mathrm{KL}(q_\lambda\parallel p)$
- *forward* $\mathrm{KL}(p\|q)$ (q conservative of uncertainty and "*includes*" p)
- *reverse* $\mathrm{KL}(q\|p)$ (q underestimates uncertainty, "*excluding*" q)

In V.I: $\mathrm{KL}(q_\lambda\|p(\cdot\mid x_{1:n},y_{1:n}))=\log p(y_{1:n}\mid x_{1:n})-\mathcal{L}(q_\lambda,p;\mathcal{D}_n)$
$\mathcal{L}(q_\lambda,p;\mathcal{D}_n)=\mathbb{E}_{\theta\sim q_\lambda}[\log p(\theta,y_{1:n}\mid x_{1:n})]+H(q_\lambda)$
$=\mathbb{E}_{\theta\sim q_\lambda}[\log p(y_{1:n}\mid x_{1:n},\theta)]-\mathrm{KL}(q_\lambda\|p(\theta))\leftarrow$ prior over θ
Evidence = const. w.r.t. $q\Rightarrow\min\mathrm{KL}\equiv\max\mathrm{ELBO}$.
$\nabla_\lambda L(q_\lambda,p;\mathcal{D}_n)=\nabla_\lambda\mathbb{E}_{\theta\sim q}[\log p(y_{1:n}\mid x_{1:n},\theta)]-\nabla_\lambda\mathrm{KL}(q\|p(\cdot))$
KL and its gradient can typically be computed exactly. For ∇_λ of exp. log-likelihood: **score gradients** / **reparametrization trick**.

8.1 Reparametrization trick

Given $\varepsilon\sim\phi$ independent of λ + diff.able and invertible function g parameterized by λ. We let $\theta\doteq g(\varepsilon;\lambda)$. Then $\varepsilon=g^{-1}(\theta;\lambda)$, and $\theta\sim q(\cdot\mid\lambda)\equiv q_\lambda(\theta)=\phi(\varepsilon)\cdot|\det(D_\varepsilon g(\varepsilon;\lambda))|^{-1}$
For "nice" function f holds $\mathbb{E}_{\theta\sim q_\lambda}[f(\theta)]=\mathbb{E}_{\varepsilon\sim\phi}[f(g(\varepsilon;\lambda))]$
As ϕ doesn't depend on λ: $\nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}[f(\theta)]=\mathbb{E}_{\varepsilon\sim\phi}[\nabla_\lambda f(g(\varepsilon;\lambda))]$

8.2 Reparametrization for Gaussians

$q_\lambda(\theta)\doteq\mathcal{N}(\theta;\mu,\Sigma)\Rightarrow\theta=g(\varepsilon;\lambda)\doteq C\varepsilon+\mu$ where $CC^T=\Sigma$
$\varepsilon=g^{-1}(\theta;\lambda)=C^{-1}(\theta-\mu)$ and $\phi(\varepsilon)=q_\lambda(\theta)\cdot|\det(C)|$

8.3 Black Box Stochastic Variational Inference

$\nabla_\lambda L(q_\lambda,p;\mathcal{D}_n)=\nabla_\lambda\mathbb{E}_{\theta\sim q_\lambda}[\log p(y_{1:n}\mid x_{1:n},\theta)]-\nabla_\lambda\mathrm{KL}(q_\lambda\|p(\cdot))$
$\approx\frac{n}{m}\sum_{j=1}^m\nabla_{C,\mu}\log p\big(y_{i_j}\mid x_{i_j},C\varepsilon^{(j)}+\mu\big)-\nabla_{C,\mu}\mathrm{KL}(q_{C,\mu}\|p(\cdot))$
where $\varepsilon^{(j)}\overset{\text{iid}}{\sim}\mathcal{N}(0,I)$ and $i_j\overset{\text{iid}}{\sim}\mathrm{Unif}([n])$.

9 Bayesian Deep Learning $\theta\sim\mathcal{N}\big(0,\sigma_p^2I\big)$

Homoscedastic likelihood: $y|x,\theta\sim\mathcal{N}(f(x;\theta),\sigma_n^2)$, then
$\hat{\theta}_{\text{MAP}}=\arg\max_\theta\log p(\theta)+\sum_{i=1}^n\log p(y_i|x_i,\theta)$
$=\arg\min_\theta\frac{1}{2\sigma_p^2}\|\theta\|_2^2+\frac{1}{2\sigma_n^2}\sum_{i=1}^n\big(y_i-f(x_i;\theta)\big)^2=$ Ridge MSE
Heteroscedastic likelihood $y|x,\theta\sim\mathcal{N}\big(f_\mu(x;\theta),f_{\sigma^2}(x;\theta)\big)$
$\hat{\theta}_{\text{MAP}}=\arg\min_\theta\frac{1}{2\sigma_p^2}\|\theta\|_2^2+\frac{1}{2}\sum_{i=1}^n\Big[\log f_{\sigma^2}(x_i;\theta)+\frac{\big(y_i-f_\mu(x_i;\theta)\big)^2}{f_{\sigma^2}(x_i;\theta)}\Big]$
can attribute some loss to large variance (aleatoric uncertainty). To model epistemic uncertainty too, perform probabilistic inference:
$p(y^*|x^*,x_{1:n},y_{1:n})=\int p(y^*|x^*,\theta)p(\theta|x_{1:n},y_{1:n})d\theta$
$p(\theta|x_{1:n},y_{1:n})=\big(p(\theta)\prod_{i=1}^n p(y_i|x_i,\theta)\big)/\int p(\theta)\prod_{i=1}^n p(y_i|x_i,\theta)d\theta$
Compute variational posterior $p(\theta|x_{1:n},y_{1:n})\approx q_\lambda\doteq q(\cdot\mid\lambda)$ and
$p(y^*|x^*,x_{1:n},y_{1:n})\approx\int p(y^*|x^*,\theta)q(\theta|\lambda)d\theta=\mathbb{E}_{\theta\sim q_\lambda(\cdot)}[p(y^*|x^*,\theta)]$
$\approx\frac{1}{m}\sum_{i=1}^m\mathcal{N}\big(y^*;f_\mu(x^*;\theta^{(i)}),f_{\sigma^2}(x^*;\theta^{(i)})\big)$ for $\theta^{(i)}\overset{\text{iid}}{\sim}q_\lambda$
Can estimate: $\mathbb{E}[y^*|x^*,x_{1:n},y_{1:n}]\approx\frac{1}{m}\sum_{i=1}^m f_\mu\big(x^*;\theta^{(i)}\big)\doteq\overline{\mu}(x^*)$
$\mathrm{Var}[y^*|x^*,x_{1:n},y_{1:n}]=\mathbb{E}_\theta\big[\mathrm{Var}_{y^*}[y^*|x^*,\theta]\big]+\mathrm{Var}_\theta\big[\mathbb{E}_{y^*}[y^*|x^*,\theta]\big]$
$\approx\frac{1}{m}\sum_{i=1}^m f_{\sigma^2}\big(x^*,\theta^{(j)}\big)+\frac{1}{m}\sum_{i=1}^m\big(f_\mu\big(x^*,\theta^{(j)}\big)-\overline{\mu}(x^*)\big)^2$

9.0.1 Monte Carlo Dropout / Dropout regularization

Dropout = variational inference with variational posterior $q(\theta|\lambda)\doteq\prod_{j=1}^d q_j(\theta_j|\lambda_j)$ where $q_j(\theta_j|\lambda_j)\doteq p\delta_0(\theta_j)+(1-p)\delta_{\lambda_j}(\theta_j)$.
$p(y^*|x^*,x_{1:n},y_{1:n})\approx\mathbb{E}_{\theta\sim q_\lambda}[p(y^*|x^*,\theta)]\approx\frac{1}{m}\sum_{i=1}^m p\big(y^*|x^*,\theta^{(i)}\big)$
where $\theta^{(i)}$ is a NN with weights λ, each set to 0 with probability p.

9.0.2 Stochastic weight averaging (SWA)

$\left[\theta^{(1)},\dots,\theta^{(T)}\right]$ T iter of SGD, approx posterior $q(\theta)\sim\mathcal{N}(\mu_{1:d}, I\sigma^2_{1:d})$
$\mu_i = 1/T\sum_{j=1}^{T}\theta_i^{(j)}$, and $\sigma_i^2 = 1/T\sum_{j=1}^{T}(\theta_i^{(j)}-\mu_i)^2$

9.0.3 Probabilistic ensembles

Sample m D_j uniformly from D, for each obtain MAP estimate $\theta^{(j)}$, approximate posterior: $p(y^*|x^*, x_{1:n}, y_{1:n}) \approx \sum_{i=1}^{m} p(y^*|x^*,\theta^{(i)})$

9.1 Calibration

Split $D_{\text{val}} = B_1 \cup \dots \cup B_m$ s.t. $\hat{p}_i \in [(j-1)/m, j/m] \forall i \in B_j$.
$\text{Acc}_j = 1/|B_j| \sum_{i\in B_j} \mathbb{1}\{\hat{y}_i = y_i\}$, $\text{Conf}_j = 1/|B_j| \sum_{i\in B_j} \hat{p}_i$ OR $\rightarrow$
$\text{freq}_j = 1/|B_j| \sum_{i\in B_j} \mathbb{1}\{y_i = 1\}$, $\text{conf}_j = 1/|B_j| \sum_{i\in B_j} \hat{p}(y_i = 1|x_i)$
Calibrated $\Leftrightarrow p(\hat{y} = y|\hat{p} = p) = p$, i.e. $\text{Acc}_j \approx \text{Conf}_j$ or $\text{freq}_j \approx \text{conf}_j$.
ECE: weighted (on $|B_j|$) avg of |Acc(freq)−C(c)onf|, MCE is max.

10 Active learning: find $S \subseteq D$ maximizing info gain

$I(S) := I(f_S; y_S) = H[f_S] - H[f_S \mid y_S] = \frac{1}{2}\log\det(I + \sigma_n^{-2} K_S)$
Theorem: $I(S)$ is *monotone submodular*:
marginal gain: $\Delta_F(x \mid A) \doteq F(A \cup \{x\}) - F(A)$. Given $A \subseteq B \subseteq \mathcal{X}$
submodular: $\Delta_F(x \mid A) \geq \Delta_F(x \mid B)$ | *monotone*: $F(A) \leq F(B)$
Theorem: greedily maximizing monotone submodular F yields constant-factor approx.: $F(S_n) \geq \left(1 - \frac{1}{e}\right) \max_{S\subseteq\mathcal{X}, |S|=n} F(S)$.

10.1 Uncertainty sampling for $S_t = \{x_1, \dots, x_t\}$

$x_{t+1} \doteq \arg\max\ \Delta_I(x \mid S_t) = \arg\max\ \frac{1}{2}\log(1 + \sigma_t^2(x)/\sigma_n^2)$
where $\sigma_t^2(x)$ is predicted variance of the GP conditioned on y_{S_t}.
Homoscedastic noise $\rightarrow \arg\max \sigma_t^2(x)$ "uncertainty sampling".
Heteroscedastic noise $\rightarrow \arg\max \sigma_t^2(x)/\sigma_n^2(x)$.

10.2 Active learning for classification (BALD)

Uncertainty sampling: $x_{t+1} \doteq \arg\max_{x\in\mathcal{X}}\ H[y_x \mid x_{1:t}, y_{1:t}]$
Mutual information distinguishes between aleatoric and epistemic:
$x_{t+1} \doteq \arg\max_{x\in\mathcal{X}}\ I(y_x; \theta \mid x, x_{1:t}, y_{1:t})$
$= \arg\max_{x\in\mathcal{X}}\ H(y_x \mid x, x_{1:t}, y_{1:t}) - \mathbb{E}_{\theta|x_{1:t},y_{1:t}} H(y_x \mid \theta, x)$

1. looks for points where different models predict different labels
2. penalizes points where models are not confident

= points x where the models *disagree*. Requires approx. posterior.

11 Bayesian optimization

Find $\arg\max_{x\in D} f^*(x)$ choosing $x_1, \dots, x_T \in D$ and observing $y_t = f^*(x_t) + \varepsilon_t$. **Cumulative regret** for *time horizon* T:
$R_T \doteq \sum_{t=1}^{T}(\max_x f^*(x) - f^*(x_t)) = T\max_x f^*(x) - \sum_{t=1}^{T} f^*(x_t)$
GOAL: $\lim_{T\to\infty} R_T/T = 0 \quad\Rightarrow\quad \lim_{T\to\infty} \max_t f(x_t) = f(x^*)$

11.1 Upper confidence sampling (optimism)

Maximize *upper confidence bound* $x_t = \arg\max \mu_{t-1}(x) + \beta_t \sigma_{t-1}(x)$.
Non-convex, can use Lipschitz-optimization (low-D) or gradient ascent (high-D). β_t scales confidence bounds to enclose f^*:
Theorem if $f^* \sim \mathcal{GP}$, and we choose β_t correctly, UCB yields $R_T/T = O^*\left(\sqrt{\gamma_T/T}\right)$ where $\gamma_T = \max_{S\subseteq\mathcal{X}, |S|\leq T} I(f_S; y_S)$. This means that if γ_T is sublinear in T, we achieve sublinear regret:
Linear ker: $\gamma_T = O(d\log T)$, Gaussian kernel: $\gamma_T = O\left((\log T)^{d+1}\right)$
Matérn kernel ($\nu > \frac{1}{2}$): $\gamma_T = O\left(T^{\frac{d}{2\nu+d}}(\log T)^{\frac{2\nu}{2\nu+d}}\right)$, Independent kernel: γ_T **linear** in T. "Smoother" f = more diminishing returns.
Theorem *(frequentist confidence intervals)* assuming $f^* \in \mathcal{H}_k(\mathcal{X})$ $R_T/T = O^*\left(\sqrt{(\beta_T\gamma_T)/T}\right)$ where $\beta_T = O\left(\|f^*\|_k + \gamma_T \log^3 T\right)$
β_t depends on γ_T and on "complexity" of f^* (its norm in $\mathcal{H}_k(\mathcal{X})$).

11.2 Thompson sampling

$x_{t+1} \in \arg\max_{x\in D} \tilde{f}(x)$, where $\tilde{f} \sim p(f \mid x_{1:t}, y_{1:t})$ sampled.

12 Diffusion Generative Models

Markov chain is a *stochastic process* with: • a prior $p(X_1)$
• *transition probabilities* $p(X_{t+1}|X_t)$ with **Markov property**:
$X_{t+1} \perp X_{1:t-1} | X_t \iff p(X_{1:T}) = p(X_1)\prod_{t=1}^{T} p(X_{t+1} \mid X_t)$
Under conditions like *ergodicity*, over time the probability of being in state x_D becomes constant and equals the stationary distribution p_∞: $X_T \to X_\infty$, *i.e.* $\lim_{t\to\infty} p(X_t = x) = p_\infty(X' = x)$.

12.1 Forward (noising) process

x_0 = data, we want x_T to be Gaussian noise. Forward process is:
$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$, $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, \beta_t I)$,
i.e. $x_t \leftarrow \sqrt{\alpha_t}x_{t-1} + \varepsilon_t$ where $\alpha_t = 1 - \beta_t$, $\varepsilon_t \sim \mathcal{N}(0, \beta_t I)$.
t-step marginal: $q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha}_t}x_0, (1 - \overline{\alpha}_t)I\right)$
where $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Schedule $\beta_t \in (0, 1)$ to be monotonically increasing, implying $\overline{\alpha}_t \to 0$ and $X_T \to \mathcal{N}(0, I)$ for $T \to \infty$.

12.2 Backward (denoising) process

A parameterized MC from $p_\lambda(X_T) = \mathcal{N}(0, I)$ to data. Notation:
• *prior* $p_\lambda(x_{1:T})$ • *likelihood* $p_\lambda(x_0|x_{1:T})$ • *posterior* $p_\lambda(x_{1:T}|x_0)$
by Markov property: $p_\lambda(x_{0:T}) = p_\lambda(x_T)\prod_{t=1}^{T} p_\lambda(x_{t-1}|x_t)$

transition prob: $p_\lambda(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\lambda(x_t, t), \Sigma_\lambda(x_t, t))$
Want to maximize *data evidence* $\log p_\lambda(x_0) = \log \int p_\lambda(x_{0:T}) dx_{1:T}$
ELBO: can be shown that for distributions $p_\lambda(x, z)$ and $q_\theta(x, z)$:

$$\log p_\lambda(x) = \mathcal{L}(\theta, \lambda; x) + \text{KL}(q_\theta(\cdot\,|x)\|p_\lambda(\cdot\,|x))$$
$$\mathcal{L}(\theta, \lambda; x) = \mathbb{E}_{z\sim q_\theta(\cdot|x)}[\log p_\lambda(x, z) - \log q_\theta(z|x)]$$

$x \equiv x_0$, $z \equiv x_{1:T}$, p_λ is the *backward m.* $q_\theta \equiv q$ is the *forward*.
Then: $\mathcal{L}(\lambda; x_0) = \mathbb{E}_{x_{1:T}\sim q(\cdot|x_0)}[\log p_\lambda(x_{0:T}) - \log q(x_{1:T}|x_0)] =$
$\underset{\text{w.r.t. }\lambda}{\text{const}} - \sum_{t=2}^{T} \mathbb{E}_{x_{1:T}\sim q(\cdot|x_0)}\left[\log \frac{q(x_{t-1}|x_t, x_0)}{p_\lambda(x_{t-1}|x_t)} + \log p_\lambda(x_0|x_1)\right] =$
$c - \sum_{t=2}^{T} \mathbb{E}_{x_t\sim q(\cdot|x_0)}[\text{KL}(q(\cdot|x_t, x_0)\|p_\lambda(\cdot|x_t))] + \mathbb{E}_{x_1\sim q(\cdot|x_0)}[\log p_\lambda(x_0|x_1)]$
Maxim. the ELBO ≡ minim. terms L_t, i.e. forcing the backward process $p_\lambda(x_{t-1}|x_t)$ to match reverse forward process $q(\cdot\,|x_t, x_0)$.
Let $L_t(x_0) = \mathbb{E}_{x_t\sim q(\cdot|x_0)}[\text{KL}(q(\cdot\,|x_t, x_0)\|p_\lambda(\cdot\,|x_t))]$:
$q(x_{t-1}|x_t, x_0) = \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t(x_0, x_0)I\right)$ for some μ_t, β_t
$p_\lambda(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\lambda(x_t, t), \Sigma_\lambda(x_t, t))$ is also gaussian:
$\text{KL}(\mathcal{N}(\mu_q, \sigma_q^2 I)\|\mathcal{N}(\mu_p, \sigma_p^2 I)) \propto 1/(2\sigma_p^2)\|\mu_p - \mu_q\|^2 + \text{const}$
Set constant variance schedule $\Sigma_\lambda(x_t, t) = \sigma_t^2 I$ and get:
$L_t(x_0) = \mathbb{E}_{x_t\sim q(\cdot|x_0)}\left[\frac{1}{2\sigma_t^2}\|\mu_\lambda(x_t, t) - \tilde{\mu}_t(x_t, x_0)\|^2 + \text{const}\right]$, where $\mu_\lambda(x_t, t)$ is learned approximator of reverse forward process mean
$\tilde{\mu}_t(x_t, x_0) = x_0(\sqrt{\overline{\alpha}_t}\beta_t)/(1 - \overline{\alpha}_t) + x_t\left(\sqrt{\alpha_t}(1 - \overline{\alpha}_{t-1})\right)/(1 - \overline{\alpha}_t)$
Non-stationary, learn noise instead, as $x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1 - \overline{\alpha}_t}\varepsilon$ where $\varepsilon \sim \mathcal{N}(0, I)$. Then, can learn to predict $\varepsilon_\lambda(x_t, t)$ instead:
$L_t(x_0) = \frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\overline{\alpha}_t)}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\left\|\varepsilon - \varepsilon_\lambda\left(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1 - \overline{\alpha}_t}, t\right)\right\|$.
Can be optimized via SGD by computing the gradient w.r.t λ.
Sampling: $x_{t-1} = 1/\sqrt{\alpha_t}\left(x_t - (1 - \alpha_t)/\left(\sqrt{1 - \overline{\alpha}_t}\right)\varepsilon_\lambda(x_t, t)\right) + \sigma_t z$

12.3 Uncertainty quantification

Instead of point estimate $\lambda = \text{MLE}(D)$ set: • prior $p(\lambda)$
• approx posterior $q(\lambda|D) \approx p(\lambda|D)$ • predictive posterior $p(x_0|D) \approx \int p(x_0|\lambda)q(\lambda|D)d\lambda$. Then filter hallucinations by rejecting points with high uncertainty $H(p(\cdot\,|D))$.

13 Markov Decision Processes

r induces sequence $R_t \doteq r(X_t, A_t)$, we usually maximize *infinite horizon discounted payoff from time* t: $G_t \doteq \sum_{m=0}^{\infty} \gamma^m R_{t+m}$
$\pi(a \mid x) \doteq \mathbb{P}(A_t = a \mid X_t = x)$ induces a MC with transitions:
$p^\pi(x'|x) \doteq \mathbb{P}(X_{t+1}^\pi = x' | X_t^\pi = x) = \sum_{a\in A} \pi(a|x)p(x'|x, a)$
$J_t(\pi) = \mathbb{E}_\pi[G_t] = \mathbb{E}[r(X_t, \pi(X_t)) + \gamma r(X_{t+1}, \pi(X_{t+1})) + \dots]$
$V_t^\pi(x) = J(\pi \mid X_t = x) = \mathbb{E}_\pi[G_t|X_t = x]$
$Q_t^\pi(x, a) = J(\pi|X_0 = x, A_0 = a) = r(x, a) + \gamma \sum_{x'\in X} p(x'|x, a) \cdot V_{t+1}^\pi(x')$
In *stationary* MDP, *Bellman Expectation Equation* holds:
$V^\pi(x) = r(x, \pi(x)) + \gamma \underset{x'\sim p(\cdot\,|\,x,\pi(x))}{\mathbb{E}}[V^\pi(x')]$, $Q^\pi(x) = r(x, a) + \gamma \underset{x'\sim p(\cdot\,|\,x,a)}{\mathbb{E}}[V^\pi(x')]$
Thanks to Bellman: $V^\pi = r^\pi + \gamma T^\pi V^\pi \Rightarrow V^\pi = (I - \gamma T^\pi)^{-1} r^\pi$
Can also use *contraction* ($\exists k < 1 s.t. \|f(x) - f(y)\| \leq k\|x - y\|$))
$B^\pi : \mathbb{R}^n \to \mathbb{R}^n$, $B^\pi v \doteq r^\pi + \gamma T^\pi v$ B^π, by *Banach fixed point theorem*, any contraction has a unique fixed point, which can be recovered via fixed point iteration exponentially fast.

13.1 Policy optimization in known MDP

Bellman's theorem: π^* optimal ⇔ greedy wrt. its induced value f.:
$\pi^* = \arg\max_{a\in A}\left[r(x, a) + \gamma\sum_{x'} p(x'|x, a)\cdot V^{\pi^*}(x')\right] = \arg\max_{a\in A} Q^{\pi^*}(x, a)$
$V^{\pi^*}(x) = \max_a\left[r(x, a) + \gamma\sum_{x'} p(x'|x, a)\cdot V^{\pi^*}(x')\right] = \max_{a\in A} Q^{\pi^*}(x, a)$
Policy iteration: Start with random π. At each iteration compute V^π and set $\pi \leftarrow \pi_{V^\pi}$. Guaranteed to monotonically improve: $V^{\pi_{t+1}}(x) \geq V^{\pi_t}(x) \forall x$, and either $\exists x \in X \mid V^{\pi_{t+1}}(x) \geq V^{\pi_t}(x)$ or $V^{\pi_t} \equiv V^*$. Finite MDPs: *exact optimum* π^* in $O^*\left(\frac{n^2 m}{1-\gamma}\right)$ iterations.
Value iteration: Start with $V_0(x) = \max_a r(x, a)$, at each iteration $V_t(x) = \max_a Q_t(x, a) = \max_a r(x, a) + \gamma\sum_{x'} P(x'|x, a)V_{t-1}(x')$
Stop when $\|V_t - V_{t-1}\|_\infty \leq \varepsilon$, then choose greedy policy w.r.t. V_t.
Converge polinomially only to ε-optimal policy, asymptotically to π^* as *Bellman upate* $B^* V_t = V_{t+1}$ is contraction and V^{π^*} is fixed p.

- policy iteration: requires policy *evaluation* at every iteration
- value iteration: for every state x, loop over actions a and reachable states $x' \rightarrow O(n \cdot m \cdot s)$ where s is sparsity coefficient

14 Reinforcement learning (basic algos)

$R_{\max}$ initialize $\hat{r}(x, a) = R_{\max}$, $\hat{p}(x^*|x, a) = 1$ where x^* is a "fairy-tale state" $s.t. \forall A\ \hat{p}(x^*|x^*, a) = 1$ and $\hat{r}(x^*, a) = R_{\max}$. Follow optimal policy for $\hat{r}$ and $\hat{p}$, after "enough" transitions, recompute π.
Lemma: every T timesteps, w.h.p. $R_{\max}$ either obtains near-optimal reward or visits at least one unknown state-action pair.
Theorem: with probability $\geq 1 - \delta$, $R_{\max}$ reaches an ε-optimal policy in step $\text{poly}(|X|, |A|, T, 1/\varepsilon, \log(1/\delta), R_{\max})$ steps.
TD-Learning $\hat{V}^\pi(x) \leftarrow (1 - \alpha_t)\hat{V}^\pi(x) + \alpha_t\left(r + \gamma\hat{V}^\pi(x')\right)$
α_t sf RM and all sa pairs chosen infinitely often⇒converges to V^π.
Off-p val. est. $\hat{Q}^\pi(x, a) \leftarrow (1-\alpha_t)\hat{Q}^\pi(x, a) + \alpha_t(r + \gamma\hat{Q}^\pi(x', \pi(x')))$
α_t sf RM and all sa pairs chosen infinitely often⇒converges to Q^π.
Q-learning $\hat{Q}^*(x, a) \leftarrow (1-\alpha_t)\hat{Q}^*(x, a) + \alpha_t\left(r + \gamma\max_{a'}\hat{Q}^*(x', a')\right)$
α_t sf RM and all sa pairs chosen infinitely often⇒converges to Q^*.
Optimistic Q-learning initialize $\hat{Q}^*(x, a) = \frac{R_{\max}}{1-\gamma}\prod_{t=1}^{T_{\text{init}}}(1 - \alpha_t)^{-1}$
With Pr $\geq 1-\delta$, obtains an ε-optimal policy after a number of steps $\text{poly}(|X|, |A|, 1/\varepsilon, \log(1/\delta), R_{\max})$. T_{init} has similar upper bound.
Score grad. $\nabla_\theta J^T(\theta) = \nabla_\theta \mathbb{E}_{\tau\sim\Pi_\theta}[g_{0:T}] = \mathbb{E}_{\tau\sim\Pi_\theta}[g_{0:T}\nabla_\theta \log\Pi_\theta(\tau)]$
W/ s.dependent baselines: $= \mathbb{E}_\tau[(g_{0:T} - b(\tau_{0:t-1}))\nabla_\theta \log\Pi_\theta(\tau)]$
ON/OFF policy **Model based:** • V/P iteration • $R_{\max}$ • Model Predictive Control (RH/MPC) • Random Shooting Methods • PETS (Probabilistic Ensembles Trajectory Sampling) • LAMBDA • H-UCRL • PBVI/PBPI (Point Based V/P Iteration) **Model free:** • TD-Learning • (Optimistic) Q-Learning • DQN (Deep Q-Network) • Double DQN • REINFORCE • Actor-Critic (A2C, A3C) • GAE/GAAC • TRPO (Trust Region Policy Optimization) • PPO (Proximal Policy Optimization) • DDPG (Deep Deterministic Policy Gradients) • TD3 (Twin Delayed DDPG) • SAC (Soft Actor-Critic) • MPO (Maximum a posteriori Policy Optimisation) • SVG (Stochastic Value Gradients) • RLHF (RL from Human Feedback)