Week 11a Sampling based logistic regression predictions

Importance sampling: a trick,

when cannot sample from p(w|D), is strange form
write integral as expectation under simple proposal distribution **q(w)**,
Monto Carlo estimate of expectation, sample from q(w) [w^(s) ~ q(w)]

$$
\begin{aligned}
P(y=1 \mid \mathbf{x}, \mathcal{D}) &= \int \sigma(\mathbf{w}^\top \mathbf{x})\, p(\mathbf{w} \mid \mathcal{D})\, \frac{q(\mathbf{w})}{q(\mathbf{w})}\, \mathrm{d}\mathbf{w} \\
&= \mathbb{E}_{q(\mathbf{w})} \left[\sigma(\mathbf{w}^\top \mathbf{x})\, \frac{p(\mathbf{w} \mid \mathcal{D})}{q(\mathbf{w})} \right] \\
&\approx \frac{1}{S} \sum_{s=1}^{S} \sigma(\mathbf{w}^{(s)\top} \mathbf{x})\, \frac{p(\mathbf{w}^{(s)} \mid \mathcal{D})}{q(\mathbf{w}^{(s)})}, \quad \mathbf{w}^{(s)} \sim q(\mathbf{w})
\end{aligned}
$$

Importance weight r^(s),

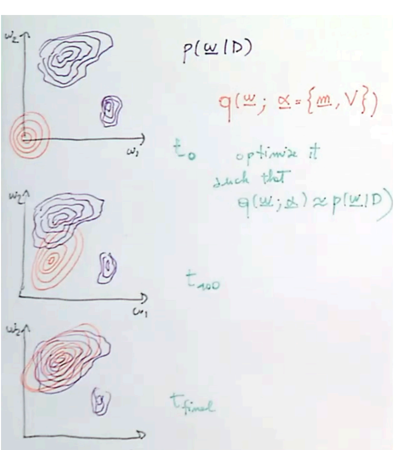
Interpretation: if p=q, weight r=1; if q is different from p, re-weighting samples, give samples more or less importance

$$
r^{(s)} = \frac{p(\mathbf{w}^{(s)} \mid \mathcal{D})}{q(\mathbf{w}^{(s)})}
$$

How to choose q?

want q(w) similar to p(w|D) (if not similar, have much less informative on importance samples)
q(w) != 0 when p(w|D) != 0, since shouldn't divide by 0
want q(w) easy to sample from

choose q from prior knowledge about shape of posterior
formalize the criteria of [q(w) similar to p(w|D)] better, systematically find out the distribution that approximates posterior (similar to Laplace approximation)
Or use iterative method, optimize divergence between p and q:



How to calculate importance weight by known terms (prior & likelihood)?
Could **approximate P(D)** using importance sampling
=>Goal: to get posterior p(w|D) above (p(w,D) known)
sample w^(s) from proposal q(w)

$$\begin{aligned}
P(\mathcal{D}) &= \int P(\mathcal{D}\,|\,\mathbf{w})\,p(\mathbf{w})\,\mathrm{d}\mathbf{w} \\
&= \int P(\mathcal{D}\,|\,\mathbf{w})\,p(\mathbf{w})\,\frac{q(\mathbf{w})}{q(\mathbf{w})}\,\mathrm{d}\mathbf{w} \\
&= \mathbb{E}_{q(\mathbf{w})}\left[\frac{P(\mathcal{D}\,|\,\mathbf{w})\,p(\mathbf{w})}{q(\mathbf{w})}\right] \\
&\approx \frac{1}{S}\sum_{s=1}^{S}\frac{P(\mathcal{D}\,|\,\mathbf{w}^{(s)})\,p(\mathbf{w}^{(s)})}{q(\mathbf{w}^{(s)})} = \frac{1}{S}\sum_{s=1}^{S}\tilde{r}^{(s)},
\end{aligned}$$

$p(w^{(s)}|D) = \frac{p(w^{(s)}, D)}{p(D)}$

IS to estimate p(D)

"unnormalized importance weights",

$$\tilde{r}^{(s)} = \frac{P(\mathcal{D}\,|\,\mathbf{w}^{(s)})\,p(\mathbf{w}^{(s)})}{q(\mathbf{w}^{(s)})}.$$

Substitution: (since bay theorem, P(w|D) = P(D|w)p(w)/p(D))

$$P(y\!=\!1\,|\,\mathbf{x},\mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S}\sigma(\mathbf{w}^{(s)\top}\mathbf{x})\,\frac{\tilde{r}^{(s)}}{\frac{1}{S}\sum_{s'=1}^{S}\tilde{r}^{(s')}}, \quad \mathbf{w}^{(s)} \sim q(\mathbf{w}) \tag{13}$$

or

$$P(y\!=\!1\,|\,\mathbf{x},\mathcal{D}) \approx \sum_{s=1}^{S}\sigma(\mathbf{w}^{(s)\top}\mathbf{x})\,r^{(s)}, \quad \mathbf{w}^{(s)} \sim q(\mathbf{w}). \tag{14}$$

In this final form, the average is under the distribution defined by the 'normalized importance weights':

$$\underline{r^{(s)}} = \frac{\tilde{r}^{(s)}}{\sum_{s'=1}^{S}\tilde{r}^{(s')}}. \tag{15}$$

++ understand & prior!!!

Week11b KL Divergence

Variational methods: another way to fit an approx. to posterior
by reducing posterior approx. problem to **optimization problem (with SGD)**
with convenient distribution **q(w; α),** over the weight w, with parameter **α**
(q could be gaussian or NN)
Set up optimization problem (**fit α**, **= mean & cov** if q is gaussian), need a cost function

(Laplace approx. = special case of variational method: define cost function, just care about mode & curvature)

KL Cost function: Measure difference between distributions, posterior p(w|D) and q(w)

Gibbs' inequality: KL >= 0; (when KL=0, p=q)
not symmetric, not satisfy triangular property (not a distance)

$$D_{\mathrm{KL}}(p \,||\, q) = \int p(\mathbf{z}) \log \frac{p(\mathbf{z})}{q(\mathbf{z})} \, \mathrm{d}\mathbf{z}.$$

But minimize D_kl (p || q) is hard & not sensible when true distribution is bimodal (use one gaussian to match it) (++why!)

So: Minimize D_KL (q || p) = variational inference objective
pick q from certain family of distribution Q (gaussian)
expand the terms:

$$D_{\mathrm{KL}}(q(\mathbf{w};\alpha) \,||\, p(\mathbf{w}\,|\,\mathcal{D})) = \int q(\mathbf{w};\alpha) \log \frac{q(\mathbf{w};\alpha)}{p(\mathbf{w}\,|\,\mathcal{D})} \, \mathrm{d}\mathbf{w} \quad = \mathbb{E}_{q(z)}\left[\log \frac{q(z)}{p(z)}\right]$$

$$= -\int q(\mathbf{w};\alpha) \log p(\mathbf{w}\,|\,\mathcal{D}) \, \mathrm{d}\mathbf{w} + \underbrace{\int q(\mathbf{w};\alpha) \log q(\mathbf{w};\alpha) \, \mathrm{d}\mathbf{w}}_{\text{negative entropy, } -H(q)}$$

First term: cross entropy between q and p (measure center of distribution q)

q is big when p is big: when approximation matches data
if q is big when p is tiny, log p close to -infinity, get very big positive penalty, when considering weights that are not compatible with the data

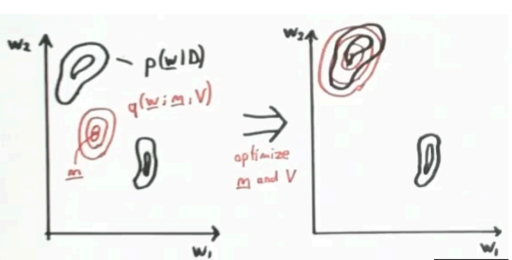
Second term: negative Entropy, (measures how spread out distribution q is)

want it small, as spread out as possible; BUT don't cover low probability regions, or the first term would grow large

=> Minimize KL change mean and cov in q(w; m, V)

approximation finds a mode of the distribution, and spread out
avoid putting mean in empty area between the mode because this the case when q is big, p is tiny (first term)

(red = approx. black = posterior)

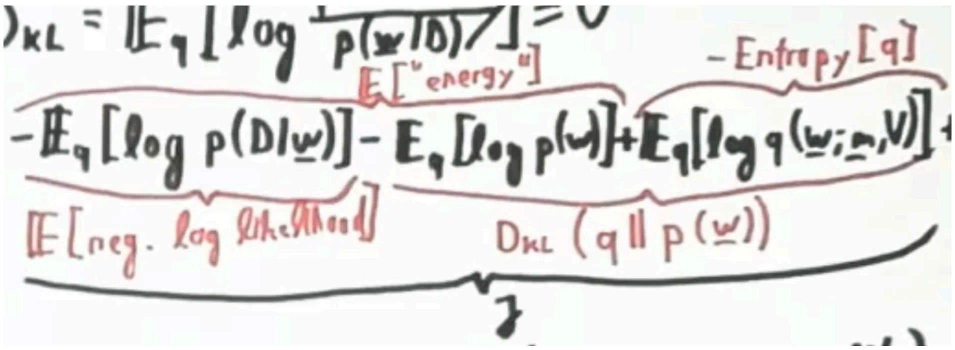


substitute posterior p(w|D) from Bayes rule:

since marginal likelihood log P(D) does not contain q (or its parameters w), independent of w, don't need to write it as expectation

$$D_{\mathrm{KL}}(q \,||\, p) = \underbrace{\mathbb{E}_q[\log q(\mathbf{w})] - \mathbb{E}_q[\log p(\mathcal{D} \,|\, \mathbf{w})] - \mathbb{E}_q[\log p(\mathbf{w})]}_{J(q)} + \log p(\mathcal{D}).$$

Many interpretations for first 3 terms (notice order is different):



Bound on marginal likelihood (useful since P(D) is difficult to calculate, could at least say it is bounded)

since would minimize J w.r.t hyper parameters (not just w.r.t mean & cov)

-J(q) = "Evidence Lower Bound (ELBO)"

$$D_{\mathrm{KL}}(q \,||\, p) \geq 0 \;\Rightarrow\; \log p(\mathcal{D}) \geq -J(q)$$

Week11c stochastic variational inference (SVI), optimize KL

have hyper parameters (e.g. prior variance σ^2), want to maximize marginal likelihood p(D | M) with respect to any hyperparameters

=> jointly **minimize J w.r.t parameters {mean, cov} and hyper parameters**

Trick 1 to make SGD works:

re-parameterized, unconstrained -> constrained

Since σ_w >0, set σ_w = e^a, optimize **a**

Since cov V symmetric, positively defined,

set V = L.L^T, (L is lower triangular matrix),

have any matrix **L tilda**

to guarantee pos def: if element Lij is diagonal, or lower triangle…

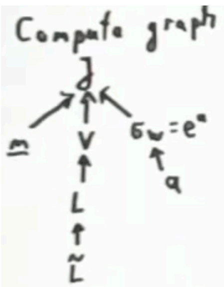
$$V = LL^T$$

$$L_{ij} = \begin{cases} e^{\tilde{L}_{ii}} & i = j \\ \tilde{L}_{ij} & i > j \\ 0 & i < j \end{cases}$$

Compute graph (un back propagation on)

IF, could calculate J from **a, L Tilda**, could BP, then could do SGD on all these parameters: m, L Tilda , a

after many SGDs, once have **L Tilda , a**, could use them to calculate **V and σw**



Evaluate cost J:

when p(D|w) is gaussian,

those 2 terms: - E_q [log p(w)] + E_q [log q(w)] is KL of 2 gaussians

could solved numerically, look up in matrix cookbook

when cannot compute likelihood p(D|w) in closed form:

Trick 2: Obtain gradient by **reparameterization trick**

since to sample a random weight w from the variational posterior,

=> sample a vector of standard normals v~N (0, I) and transform it: w = m + Lv

f could be log likelihood log p(D|w), or general function

L builds cov V, this transformation will yield a gaussian with final cov V

when integrate (expectation) over v, does not depends on w (integrate over in J)

$$\mathbb{E}_{\mathcal{N}(\mathbf{w};\mathbf{m},V)}[f(\mathbf{w})] = \mathbb{E}_{\mathcal{N}(\boldsymbol{\nu};\mathbf{0},\mathbb{I})}[f(\mathbf{m} + L\boldsymbol{\nu})]$$

Write down derivatives: (memorize!)

line 1: push gradient into integral sign (expectation)

line 2: one sample approximation (instead of monte carlo, for efficiency reason)

is unbiased estimator of initial gradient

$$\begin{aligned}
\nabla_{\mathbf{m}} \mathbb{E}_{\mathcal{N}(\mathbf{w};\,\mathbf{m},V)}[f(\mathbf{w})] &= \mathbb{E}_{\mathcal{N}(\boldsymbol{\nu};\,\mathbf{0},\mathbb{I})}[\nabla_{\mathbf{m}} f(\mathbf{m}+L\boldsymbol{\nu})] \\
&\approx \nabla_{\mathbf{m}} f(\mathbf{m}+L\boldsymbol{\nu}), \quad \boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0},\mathbb{I}), \\
\nabla_{L} \mathbb{E}_{\mathcal{N}(\mathbf{w};\,\mathbf{m},V)}[f(\mathbf{w})] &\approx \nabla_{L} f(\mathbf{m}+L\boldsymbol{\nu}), \quad \boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0},\mathbb{I}) \\
&= [\nabla_{\mathbf{w}} f(\mathbf{w})]\boldsymbol{\nu}^{\top}, \quad \mathbf{w} = \mathbf{m}+L\boldsymbol{\nu}, \;\; \boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0},\mathbb{I})
\end{aligned}$$