120??

DOI: *The Canadian Journal of Statistics / La revue canadienne de statistique*

Theorem

[the-

o-

rem]Corollary

[the-

o-

rem]Definition

[theorem]Lemma

[the-

o-

rem]Proposition

[the-

o-

rem]Remark

The Canadian Journal of Statistics

3Vol. xx, No. yy, 20??, Pages 1–25

La revue canadienne de statistique

Weighted Bayesian Bootstrap for Scalable Bayes

BlindedA[1*]

BlindedB[2]

[1]*Author affiliations will go*

© 20?? Statistical Society of Canada / Société statistique du Canada

here in the accepted manuscript, but do NOT include them in your initial submission because it must

be

anonymous.

[2]*Second*

Affiliation

Bayesian,

Boot-

strap,

MCMC,

Weighted

Boot-

strap,

ABC,

Trend

Fil-

ter-

ing,

Deep

Learn-

ing,

Ten-

sor-

Flow,

Reg-

u-

lar-

iza-

tion:

MSC

2010:

Pri-

mary

62???;

sec-

ondary

62???

We

de-

velop

a

weighted

Bayesian

Boot-

strap

(WBB)

for

ma-

chine

learn-
ing
and
statis-
tics.
WBB
pro-
vides
un-
cer-
tainty
quan-
tifi-
ca-
tion
by
sam-
pling
from
a
high
di-
men-
sional
pos-

terior distribution. WBB is computationally fast and scalable using only off-the-shelf opti-

miza-
tion
soft-
ware
such
as
Ten-
sor-
Flow.
We
pro-
vide
reg-
u-
lar-
ity
con-
di-
tions
which
ap-
ply
to
a
wide

range of machine learning and statistical models. We illustrate our methodology in regularized re- gres- sion, trend fil- ter- ing and deep learn- ing. Fi- nally, we con- clude with di- rec- tions for fu- ture research.

The Canadian Journal of Statistics xx: 1–25; 20??
© 20?? Statistical Society of Canada

Résumé: Insérer votre résumé ici. We will

supply a French abstract for those authors who can't prepare it themselves. *La revue canadienne de statistique* xx: 1–

25; 20??

© 20?? Société statistique du Canada

* *Author to whom correspondence may be addressed. E-mail: Insert your email address here only after your paper has been accepted*

1. INTRODUCTION

Weighted Bayesian Bootstrap (WBB) is a simulation-based algorithm for assessing uncertainty in machine learning and statistics. Uncertainty quantification (UQ) is an active area of re- search, par- tic- u- larly in high- dimensional in- fer- ence prob- lems. Whilst there are com- pu- ta-

tion-
ally
fast
and
scal-
able
al-
go-
rithms
for
train-
ing
mod-
els
in
a
wide
va-
ri-
ety
of

con-
texts,
un-
cer-
tainty
as-
sess-
ment
is
still
re-
quired.
De-
vel-
op-
ing
com-
pu-
ta-
tion-
ally
fast

scalable algorithms for sampling a posterior distribution is a notoriously hard prob- lem. WBB makes a con- tri- bu- tion to this lit- er- a- ture by show- ing how off-

the-

shelf

op-

ti-

miza-

tion

al-

go-

rithms,

such

as

con-

vex

op-

ti-

miza-

tion

or

stochas-

tic

gra-
di-
ent
de-
scent
(SGD)
in
Ten-
sor-
Flow
can
also
be
used
to
pro-
vide
un-
cer-
tainty
quan-
tifi-

ca-

tion.

Our

work

builds

on

?

who

pro-

vide

a

weighted

like-

li-

hood

Boot-

strap

(WLB)

method

for

Bayesian

in-
fer-
ence.
They
de-
velop
a
weighted
like-
li-
hood
Boot-
strap
al-
go-
rithm
to-
gether
with
the
ap-

pro-
pri-
ate
asymp-
totic
anal-
y-
sis
to
show
that
such
an
al-
go-
rithm
pro-
vides
ef-
fi-
cient
pos-

te-
rior
sam-
ples.
Their
boot-
strap
pro-
ce-
dure
ex-
ploits
the
fact
that
the
pos-
te-
rior
dis-
tri-

bu-
tion
cen-
tered
at
the
max-
i-
mum
like-
li-
hood
es-
ti-
mate
(MLE)
has
a
sec-
ond
or-

der ex- pan- sion that also de- pends on the prior and its deriva- tive. The weighted Bayesian Boot- strap (WBB) calculates a series of posterior modes rather than MLEs. This has the advantage that high di- men- sional pos- te- rior modes are read- ily avail- able par- tic- u- larly us- ing the regu- lar- ized es- ti- mates are fast to com- pute from con- vex op- ti- miza- tion meth- ods or stochastic gra- di- ent de- scent (SGD) for neu- ral net- work ar- chi- tec- tures such as deep learn- ing.

By linking WLB and WBB, with modern-day optimization to calibrate estimate, we provide a framework for uncertainty quantification.

Uncertainty estimates are provided at

little to no extra cost. Quantifying uncertainty is typically unavailable in a purely reg- u- lar- iza- tion op- ti- miza- tion method. An- other fea- ture that is straight- forward to add is a reg- u- lar- iza- tion path across hyper- parameters. This is so much eas- ier than

tra-
di-
tional
Bayesian
to
do
prior
sen-
si-
tiv-
ity
anal-
y-
sis
where
hyper-
parameters
are
hard
to
as-

sess. Rather we use predictive cross-validation techniques.

The rest of the paper is outlined as

fol- lows. Sec- tion 2 de- vel- ops our weighted Bayesian Boot- strap (WBB) al- go- rithm. Sec- tion 3 provides an ap- pli- ca- tion to high di- men- sional sparse re- gres- sion, trend fil- ter- ing and deep learn- ing. WBB can also be ap- plied to Bayesian tree mod- els (**?**). Fi- nally, Sec- tion 4 con- cludes with directions for future research. Areas for future study include Bootstrap filters

in
state-
space
mod-
els
(**?**)
and
com-
par-
i-
son
with
the
resampling-
sampling
per-
spec-
tive
to
se-
quen-

tial Bayesian inference (**?**), etc.

2. WEIGHTED BAYESIAN BOOTSTRAP

Let y an n-vector of outcomes, θ

de-

notes

a

d-

dimensional

pa-

ram-

e-

ter

of

in-

ter-

est

and

A

a

fixed

$n \times$

d

ma-

trix

whose rows are the design points (or "features") a_i^T where we index observations by i

and pa- ram- e- ters by j. A large num- ber of ma- chine learn- ing and sta- tis- ti- cal problems can be expressed in the form

$$\underset{\theta \in \mathcal{R}^d}{\text{minimize}} \quad l(y|\theta) + \lambda \phi(\theta)\,,$$

where $l(y|\theta) = \sum_{i=1}^{n} \log f(y_i; a_i^\top \theta)$ is a measure of fit (or “em- pir- i- cal risk func- tion”) de- pend- ing im- plic- itly on A and y. The penalty func- tion or reg- u- lar- iza- tion term, $\lambda\phi(\theta)$, ef- fects a fa- vor- able bias- variance trade- off. We al- low for the

pos-

si-

bil-

ity

that

$\phi(\theta)$

may

have

points

in

its

do-

main

where

it

fails

to

be

dif-

fer-

en-
tiable.

Suppose that we ob- serve data, $y = (y_1, \ldots, y_n)$ from a model pa- ram- e- ter- ized by θ. For

ex-
am-
ple,
we
might
have
a
prob-
a-
bilis-
tic
model
that
de-
pends
on
a
pa-
ram-
e-
ter,
θ,

where $p(y|\theta)$ is known as the likelihood function. Equivalently, we can define a measure of fit $l(y|\theta) = \log f(y;\theta) = \log p(y|\theta)$. We will make use of the following

(i) Let $\hat{\theta}_n := \mathrm{argmax}_\theta\ p(y|\theta)$ be the MLE,

(ii) Let $\theta_n^* := \operatorname{argmax}_\theta p(\theta|y)$ be the posterior mode,

(iii) Let $\bar{\theta}_n := E(\theta|y)$ be the posterior mean.

We now develop a key duality between regularization and posterior bootstrap simulation.

2.1. Bayesian Regularization Duality

From the Bayesian perspective, the

mea-
sure
of
fit,
$l(y|\theta) =$
$-\log f(y;\theta)$,
and
the
penalty
func-
tion,
$\lambda\phi(\theta)$,
cor-
re-
spond
to
the
neg-
a-
tive
log-
a-

rithms of the like-li-hood and prior dis-tri-bu-tion in the hi-er-ar-chi-cal model

$$f(y;\theta) = p(y|\theta) \propto \exp\{-l(y|\theta)\}\ , \quad p(\theta) \propto \exp\{-\lambda\phi(\theta)\}$$

$$p(\theta|y) \propto \exp\{-(l(y|\theta) + \lambda\phi(\theta))\}.$$

The prior is not nec- es- sar- ily proper but the pos- te- rior, $p(\theta|y) \propto p(y|\theta)p(\theta)$, may still be proper. This provides an equivalence between regularization and Bayesian methods.

For ex- am- ple, re- gres- sion with a least squares log- likelihood sub- ject to a penalty such as an

L^2-norm (ridge) Gaussian probability model or L^1-norm (lasso) double exponential prob-

a-

bil-

ity

model.

We

then

have

$$\hat{\theta}_n = _{\theta \in \Theta} l(y|\theta), \tag{2}$$

$$\theta_n^* = _{\theta \in \Theta} \{l(y|\theta) + \lambda \phi(\theta)\} \tag{3}$$

Let

∂

be

the

sub-

d-

if-

fer-

en-

tial

operator. Then a necessary and sufficient condition for θ^* to minimize $l(y|\theta) +$

$\lambda\phi(\theta)$

is

$$0 \in \partial \left\{ l(y|\theta) + \lambda\phi(\theta) \right\} = \nabla l(y|\theta) + \lambda\partial\phi(\theta)$$

the

sum

of

a

point

and

a

set.

The

op-

ti-

miza-

tion

lit-

er-

a-

ture

characterizes θ^* as the fixed point of a proximal operator $\theta^* = \mathrm{prox}_{\gamma\phi}\{\theta^* - \lambda\nabla f(\theta^*)\}$, see **?** and **?** for further discussion.

A general class of natural exponential fam- ily mod- els can be ex- pressed in terms of the Breg- man di- ver- gence of the dual of the cu- mu- lant trans- form. Let ϕ be the con- ju- gate Leg- en- dre trans- form of ψ.

Hence $\psi(\theta) = \sup_\mu \left(\mu^\top \theta - \phi(\mu)\right)$. Then we can write

$$
\begin{aligned}
p_\psi(y|\theta) &= \exp\left(y^\top \theta - \psi(\theta) - h_\psi(y)\right) \\
&= \exp\left\{\inf_\mu \left((y-\mu)^\top \theta - \phi(\mu)\right) - h_\psi(y)\right\} \\
&= \exp\left(-D_\phi(y, \mu(\theta)) - h_\phi(y)\right)
\end{aligned}
$$

where the infimum is attained at

$\mu(\theta) =$
$\phi'(\theta)$
is
the
mean
of
the
ex-
po-
nen-
tial
fam-
ily
dis-
tri-
bu-
tion.
We
rewrite
$h_\psi(y)$
is

terms of the correction term and $h_\phi(y)$. Here there is a duality as D_ϕ can be interpreted as a Bregman divergence.

For a wide range of nonsmooth objective functions/statistical models, recent regularization methods provide fast, scalable algorithms for

cal-
cu-
lat-
ing
es-
ti-
mates
of
the
form
(**??**),
which
can
also
be
viewed
as
the
pos-
te-
rior

mode. Therefore as λ varies we obtain a full regularization path as a form of prior sen- si- tiv- ity anal- y- sis.

? and ? con- sid- ered sce- nar- ios where pos- terior modes can be used as posterior means from augmented probability models. Moreover, in their orig- i- nal foun- da- tion of the Weighted Like- li- hood Boot- strap (WLB), **?** in- tro- duced the concept of the implicit prior. Clearly this is an avenue for future research.

2.2. WBB Algorithm

We now define the weighted Bayesian Bootstrap (WBB). Following **?**, we construct a randomly weighted posterior distribution denoted by

$$\mathbf{w} = (w_1, ..., w_n, w_p),\ p_{\mathbf{w}}(\theta|y) \propto \prod_{i=1}^{n} p(y_i|\theta)^{w_i} p(\theta)^{w_p}$$

where the weights $w_p, w_i \sim Exp(1)$

are randomly generated weights. It's equivalent to draw $w_i = \log(1/U_i)$ where U_i's are i.i.d. Uniform (0,1), which is mo- ti- vated by the uni- form Dirich- let dis- tri- bu- tion for multi- no- mial data. We have used the fact that for i.i.d. ob- ser- va- tions, the like- li- hood can be fac- tor- ized as

$p(y|\theta) =$

$\prod_{i=1}^{n} p(y_i|\theta)$.

This

is

not

cru-

cial

for

our

anal-

y-

sis

but

is

a

com-

mon

as-

sump-

tion.

Let

$\theta^*_{\mathbf{w},n}$ denote the mode of this regularized distribution. Again, there is an equivalence

$$\theta^*_{\mathbf{w},n} := {}_{\theta}\, p_{\mathbf{w}}(\theta|y) \equiv {}_{\theta} \sum_{i=1}^{n} w_i l_i(y_i|\theta) + \lambda w_p \phi(\theta)$$

where $l_i(y_i|\theta) = -\log p(y_i|\theta)$ and $\lambda\phi(\theta) = -\log p(\theta)$. Note that we have a weighted like-li-hood and a new reg- u- lar- iza- tion pa- ram- e- ter, λw_p.

The crux of our pro- ce- dure is to create a sample of the weighted posterior modes $\{\theta^*_{\mathbf{w},n}\}$ (computationally cheap as each sub-problem can be solved via op- ti- miza- tion). Our main re- sult is the fol- low- ing:

Algorithm: Weighted

Bayesian Bootstrap (WBB)

1. Iterate: sample $\mathbf{w} = \{w_1, w_2, ..., w_n, w_p\}$ via exponentials. $w_p, w_i \sim Exp(1)$.
2. For each $\mathbf{w}$,

solve

$$\theta^*_{\mathbf{w},n} = {}_{\theta} \sum_{i=1}^{n} w_i l_i(\theta) + \lambda w_p \phi(\theta).$$

The WBB algorithm is fast and scalable to compute a regularized estimator. For a large number of popular priors, the minimizing so-

lu-

tion

$\theta^*_{\mathbf{w},n}$

in

the

sec-

ond

step

can

be

di-

rectly

ob-

tained

via

reg-

u-

lar-

iza-

tion

packages such as `glmnet` by Trevor Hastie and `genlasso` by Taylor Arnold. When the likelihood function or the prior is spe- cially de- signed, Stochas- tic Gra- di- ent De- scent (SGD) is pow- er- ful and fast enough to solve the min-i-miza-tion prob-lem. It can be eas-ily im-ple-mented in Ten-sor-Flow once the objective function is specified. See Appendix (**??**) and **?** for further discussion.

The next section builds on **?** and derives asymptotic properties of the weighted Bayesian

Boot-
strap.
We
sim-
ply
add
the
reg-
u-
lar-
ized
fac-
tor.
To
choose
the
amount
of
reg-
u-
lar-

iza-
tion
λ,
we
can
use
the
marginal
like-
li-
hood
$m_\lambda(y)$,
es-
ti-
mated
by
bridge
sam-
pling
(**?**)
or

simply using predictive cross-validation.

2.3. WBB Properties

The following proposition

which follows from the Theorem 2 in **?** summaries the properties of WBB.

Proposition

The weighted Bayesian Bootstrap draws are approximate posterior samples

$$\left\{\theta^{*(k)}_{\mathbf{w},n}\right\}_{k=1}^{K} \sim p(\theta|y).$$

Now we consider 'large n' properties. The variation in the posterior density $p(\theta|y) \propto$

$e^{-nl_n(\theta)}p(\theta)$

for

suf-

fi-

ciently

large

n

will

be

dom-

i-

nated

by

the

like-

li-

hood

term.

Ex-

pand-

ing

$l_n(\theta)$ around its maximum, $\hat{\theta}$, and defining $J_n(\hat{\theta}) = nj(\hat{\theta})$ as the observed information matrix gives the tra- di- tional nor- mal ap- prox- i- ma- tion for the pos- te- rior dis- tri- bu-tion

$$\theta \sim N_d\left(\hat{\theta}_n, J_n^{-1}(\hat{\theta})\right)$$

where $\hat{\theta}_n$ is the MLE. A more accurate approximation is obtained by ex- pand- ing around the pos- te- rior mode, θ^*, which we will ex- ploit in our weighted Bayesian Bootstrap. Now we have the asymptotic distributional approximation

$$\theta \sim N_d\left(\theta^*, J_n^{-1}(\theta^*)\right)$$

where $\theta_n^* :=$ arg $\max_\theta\, p(\theta|y)$

is the pos- te- rior mode.

The use of the pos- te- rior mode here is cru- cially im- portant as it’s the mode that is computationally available from TensorFlow and Keras. Ap-

prox-
i-
mate
nor-
mal-
ity
and
sec-
ond
or-
der
ap-
prox-
i-
ma-
tion
also
holds,
see
?,
?
and

?

for future discussion. Specifically,

$$\sqrt{nI(\hat{\theta}_n)}\left(\theta_n^* - \hat{\theta}_n\right) \stackrel{D}{=} Z$$

where $Z \sim N(0, 1)$ is a standard Normal variable. The conditional posterior satisfies $\mathbb{P}\left(|\theta_n^* - \hat{\theta}_n| > \epsilon\right) \to 0$ for each $\epsilon > 0$ as $n \to \infty$.

In the ’large p’ case, a number of results are available for posterior concentration, for ex- am- ple, see **?** for sparse high di- men- sional mod- els.

3. APPLICATIONS

Consider now a number of scenarios to assess when WBB corresponds to a full Bayesian pos- te- rior dis- tri- bu- tion.

3.1. Lasso

First, a sim- ple uni- vari- ate normal means problem with a lasso prior where

$$y|\theta \sim N(\theta, 1^2), \quad \theta \sim Laplace(0, 1/\lambda)$$

Given the i.i.d. exponential weights w_1 and

w_2, the weighted posterior mode $\theta^*_{\mathbf{w}}$ is

$$\theta^*_{\mathbf{w}} = {}_{\theta \in \Theta} \left\{ \frac{w_1}{2}(y-\theta)^2 + \lambda w_2 |\theta| \right\}.$$

This is sufficiently simple for an exact WBB solution in terms of soft thresholding:

$$\theta^*_{\mathbf{w}} = \begin{cases} y - \lambda w_2/w_1 & \text{if } y > \lambda w_2/w_1, \\ y + \lambda w_2/w_1 & \text{if } y < -\lambda w_2/w_1, \\ 0 & \text{if } |y| \leq \lambda w_2/w_1. \end{cases}$$

The WBB mean $E_{\mathbf{w}}(\theta^*_{\mathbf{w}}|y)$ is

ap-
prox-
i-
mated
by
the
sam-
ple
mean
of
$\{\theta_{\mathbf{w}}^{*(k)}\}_{k=1}^{K}$.
On
the
other
hand,
?
gives
the
ex-
pres-
sion
for

the posterior mean,

$$E(\theta|y) = \frac{\int_{-\infty}^{\infty} \theta \exp\{-(y-\theta)^2/2 - \lambda|\theta|\}\, d\theta}{\int_{-\infty}^{\infty} \exp\{-(y-\theta)^2/2 - \lambda|\theta|\}\, d\theta}$$
$$= \frac{F(y)}{F(y)+F(-y)}(y+\lambda) + \frac{F(-y)}{F(y)+F(-y)}(y-\lambda)$$
$$= y + \frac{F(y)-F(-y)}{F(y)+F(-y)}\lambda$$

where $F(y) = \exp(y)\Phi(-y-\lambda)$ and $\Phi(\cdot)$ is the c.d.f. of standard normal distribution. We plot the WBB mean versus the exact posterior mean in Figure (**??**). Interestingly, WBB algorithm gives sparser posterior means. To illustrate

3.2. Diabetes Data

simple.pdf

FIGURE 1: Normal means model with lasso prior: WBB mean $E_{\mathbf{w}}(\theta^*_{\mathbf{w}}|y)$ (in solid

our method-ology, we use weighted Bayesian Boot-strap (WBB) on the clas-sic di-a-betes dataset. The mea-sure-ments for 442 diabetes patients are obtained ($n = 442$), with 10 baseline variables ($p = 10$), such as age, sex, body mass index, average blood pressure, and six blood serum measurements.

The likelihood function is given by

$$l(y|\beta) = \prod_{i=1}^{n} p(y_i|\beta)$$

where

$$p(y_i|\beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right\}.$$

We draw 1000 sets of weights

$\mathbf{w} =$

$\{w_i\}_{i=1}^{n+1}$

where

w_i's

are

i.i.d.

ex-

po-

nen-

tials.

For

each

weight

set,

the

weighted

Bayesian

es-

ti-

mate

$\beta^*_{\mathbf{w}}$

is

cal- cu- lated us- ing (**??**) via the reg- u- lar- iza- tion method in the pack- age `glmnet`.

$$\hat{\beta}_{\mathbf{w}} := {}_{\beta} \sum_{i=1}^{n} w_i(y_i - x_i'\beta)^2 + \lambda w_{n+1} \sum_{j=1}^{p} |\beta_j| \ .$$

The reg- u- lar- iza- tion fac- tor λ is cho- sen by cross- validation with un- weighted like- li- hood. The weighted Bayesian Boot- strap is also per- formed with fixed prior, namely, w_{n+1} is set to be

1 for all boot-strap sam-ples. **?** an-a-lyze the same dataset us-ing the Bayesian Bridge es-ti-

mator and suggest MCMC sampling from the posterior.

To compare our WBB results we

also run the Bayesian bridge estimation. Here the Bayesian setting we use is

$$p(\beta, \sigma^2) = p(\beta|\sigma^2)p(\sigma^2), \text{ where } p(\sigma^2) \propto 1/\sigma^2.$$

The prior

on β, with suitable normalization constant C_α, is given by

$$p(\beta) = C_\alpha \exp(-\sum_{j=1}^{p} |\beta_j/\tau|^\alpha).$$

The hyperparameter is

drawn as $\nu = \tau^{-\alpha} \sim \Gamma(2,2)$, where $\alpha = 1/2$.

Figure (??) shows the results of all these three methods (the weighted Bayesian Boot-strap with

diabetes.pdf

FIGURE 2: Diabetes example: the weighted Bayesian Bootstrap (with fixed prior and weighted prior) and Bayesian Bridge are used to draw from the marginal posteriors for β_j's, j = 1,2,...10.

fixed prior / weighted prior and the Bayesian Bridge). Marginal posteriors for β_j's are presented. One notable fea- ture is that the weighted Bayesian Boot- strap tends to in- tro- duce more spar- sity than Bayesian Bridge does. For example, the weighted Bayesian Bootstrap posteriors of `age`, `ldl` and `tch` have higher spikes located around 0, compared with the Bayesian Bridge ones. For `tc`, `hdl`, `tch` and `glu`, multi-modes in the marginal pos- te- ri- ors are ob- served. In gen- eral, the pos- te- ri- ors with fixed pri- ors are more con- cen- trated than those with ran- domly weighted pri- ors. This dif- fer- ence is nat- u- rally attributed to the certainty in the prior weights.

3.3. Trend Filtering

The generalized lasso solves the

optimization problem:

$$\beta^* = {}_{\beta} \left\{ l(y|\beta) + \lambda\phi(\beta) \right\} \quad (6)$$

$$= {}_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|D\beta\|_1 \quad (7)$$

where $l(y|\beta) = \frac{1}{2}\|y - X\beta\|_2^2$ is the negative log-likelihood. $D \in$

$\mathcal{R}^{m\times p}$ is a penalty ma- trix and $\lambda\phi(\beta) = \lambda\|D\beta\|_1$ is the neg- a- tive log- prior or reg- u- lar- ization penalty. There are fast path algorithms for solving this problem (see `genlasso` package).

As a sub-problem, polynomial trend filtering (**??**) is recently introduced for piecewise poly- no- mial curve- fitting, where the knots and the pa- ram- e- ters are cho- sen adap- tively. Intuitively, the trend-filtering estimator is similar to an adaptive spline model: it pe-

nal-
izes
the
dis-
crete
deriva-
tive
of
or-
der
k,
re-
sult-
ing
in
piece-
wise
poly-
no-
mi-
als

of higher degree for larger k.

Specifically, $X = I_p$ in the trend filtering setting and the data $y =$ $(y_1, ..., y_p)$ are as- sumed to be mean- ing- fully or- dered from 1 to p. The penalty ma- trix is specially designed by the discrete $(k+1)$-th order derivative,

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 \dots & 0 & 0 \\ 0 & -1 & 1 \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 \dots & -1 & 1 \end{bmatrix}_{(p-1)\times p}$$

and $D^{(k+1)} = D^{(1)} D^{(k)}$ for $k = 1, 2, 3....$ For example, the log-prior in linear trend filtering is explicitly written as $\lambda \sum_{i=1}^{p-2} |\beta_{i+2} - 2\beta_{i+1} + \beta_i|$. For a general order $k > 1$,

$$\|D^{(k+1)}\beta\|_1 = \sum_{i=1}^{p-k-1} \Big| \sum_{j=i}^{i+k+1} (-1)^{(j-i)} \binom{k+1}{j-i} \beta_j \Big|.$$

WBB solves

the following generalized lasso problem in each draw:

$$
\begin{aligned}
\beta^*_{\mathbf{w}} &= \underset{\beta}{} \frac{1}{2}\sum_{i=1}^{p} w_i(y_i - \beta_i)^2 + \lambda w_{p+1}\|D^{(k)}\beta\|_1 \\
&= \underset{\beta}{} \frac{1}{2}\|Wy - W\beta\|_2^2 + \lambda\|D^{(k)}\beta\|_1 \\
&= W^{-1} \underset{\tilde{\beta}}{} \frac{1}{2}\|\tilde{y}_{\mathbf{w}} - \tilde{\beta}_{\mathbf{w}}\|_2^2 + \lambda\|\tilde{D}^{(k)}_{\mathbf{w}}\tilde{\beta}_{\mathbf{w}}\|_1
\end{aligned}
$$

where

$$W = diag(\sqrt{w_i}/\sqrt{w_{p+1}}, ..., \sqrt{w_p}/\sqrt{w_{p+1}})$$

and

$$\tilde{y}_{\mathbf{w}} = Wy,\ \tilde{\beta}_{\mathbf{w}} = W\beta,\ \tilde{D}^{(k)}_{\mathbf{w}} = D^{(k)}W^{-1}.$$

To illustrate our method, we simulate data y_i from a

Fourier se- ries re- gres- sion

$$y_i = \sin\left(\frac{4\pi}{500}i\right)\exp\left(\frac{3}{500}i\right) + \epsilon_i$$

for $i =$ $1, 2, ...500$, where $\epsilon_i \sim$ $N(0, 2^2)$ are i.i.d. Gaus- sian noises. The cubic trend filtering result is given in Figure (**??**).

For each i, the weighted Bayesian Bootstrap gives a group of estimates $\{\beta^*_{\mathbf{w}}(i)\}_{j=1}^{T}$ where T is the total number of draws. The standard er- ror of $\hat{\beta}_i$ is eas- ily com- puted us- ing these weighted boot- strap es- ti- mates.

filtering.pdf

3.4. Deep Learning: MNIST Example

Deep learning is a form of machine learning that uses hi-er- ar- chi- cal ab- stract lay- ers of la- tent vari- ables to per- form pat- tern match- ing and pre-

dic-
tion.
?
take
a
Bayesian
prob-
a-
bilis-
tic
per-
spec-
tive
and
pro-
vide
a
num-
ber
of
in-

sights into more efficient algorithms for optimization and hyperparameter tuning.

The gen- eral goal is to finds a pre- dic- tor of an out- put y given a high di- mensional in- put x. For a clas- si- fi- ca- tion prob- lem, $y \in \{1, 2, ..., K\}$ is a dis- crete vari- able and can be coded as a K-dimensional 0-1 vector. The model is as follows. Let $z^{(l)}$ denote the l-th layer, and so $x = z^{(0)}$. The final output is the response y, which can be nu-meric or cat-e-gor-i-cal. A deep pre-dic-tion rule is then

$$
\begin{aligned}
z^{(1)} &= f^{(1)}\Big(W^{(0)}x + b^{(0)}\Big),\\
z^{(2)} &= f^{(2)}\Big(W^{(1)}z^{(1)} + b^{(1)}\Big),\\
&\cdots\\
z^{(L)} &= f^{(L)}\Big(W^{(L-1)}z^{(L-1)} + b^{(L-1)}\Big),\\
\hat{y}(x) &= z^{(L)}.
\end{aligned}
$$

Here, $W^{(l)}$ are weight matrices, and $b^{(l)}$ are threshold or ac-

ti-

va-

tion

lev-

els.

$f^{(l)}$

is

the

ac-

ti-

va-

tion

func-

tion.

Prob-

a-

bilis-

ti-

cally,

the

output y in a classification problem is generated by a probability model

$$p(y|x, W, b) \propto \exp\{-l(y|x, W, b)\}$$

where $l(y|x, W, b) = \sum_{i=1}^{n} l_i(y_i|x_i, W, b)$ is the neg- a- tive cross- entropy,

$$l_i(y_i|x_i, W, b) = l_i(y_i, \hat{y}(x_i)) = \sum_{k=1}^{K} y_{ik} \log \hat{y}_k(x_i)$$

where y_{ik} is 0 or 1

and $K = 10$. Adding the neg- a- tive log- prior $\lambda\phi(W, b)$, the ob- jec- tive func- tion (neg- a- tive log-posterior) to be min- i- mized by stochas- tic gra- di- ent de- scent is

$$\mathcal{L}_\lambda(y, \hat{y}) = \sum_{i=1}^{n} l_i(y_i, \hat{y}(x_i)) + \lambda\phi(W, b).$$

Accordingly, with each draw

of weights $\mathbf{w}$, WBB provides the estimates $(W^*_{\mathbf{w}}, b^*_{\mathbf{w}})$ by solving the following optimization problem.

$$(W^*_{\mathbf{w}}, b^*_{\mathbf{w}}) =_{W,b} \sum_{i=1}^{n} w_i l_i(y_i|x_i, W, b) + \lambda w_p \phi(W, b)$$

We take the classic MNIST example to illustrate the application of WBB in deep learning. The MNIST database of handwritten digits, available from Yann Le-

Cun's web- site, has 60,000 train- ing ex- am- ples and 10,000 test ex- am- ples. Here the high- dimensional x

is a normalized and centered fixed-size (28×28) image and the output $\hat{y}$ is a

10-

dimensional

vec-

tor,

where

i-

th

co-

or-

di-

nate

cor-

re-

sponds

to

the

prob-

a-

bil-

ity

of

that image being the i-th digit.

For simplicity, we build a 2-layer neural

net-
work
with
layer
sizes
128
and
64
re-
spec-
tively.
There-
fore,
the
di-
men-
sions
of
pa-
ram-
e-

ters are

$$W^{(0)} \in \mathcal{R}^{128\times 784},\ b^{(0)} \in \mathcal{R}^{128},$$
$$W^{(1)} \in \mathcal{R}^{64\times 128},\ b^{(1)} \in \mathcal{R}^{64},$$
$$W^{(2)} \in \mathcal{R}^{10\times 64},\ b^{(0)} \in \mathcal{R}^{10}.$$

The activation function $f^{(i)}$ is ReLU, $f(x) = \max\{0, x\}$, and the neg-

a-
tive
log-
prior
is
spec-
i-
fied
as

$$\lambda\phi(W,b) = \lambda\sum_{l=0}^{2}\|W^{(l)}\|_2^2$$

where
$\lambda =$
10^{-4}.

Figure
(**??**)
shows
the
pos-
te-

rior dis- tri- bu- tion of the clas- si- fi- ca- tion ac- cu- racy in the test dataset. We see that the test accuracies are centered around 0.75 and the posterior distribution is left-skewed. Furthermore, the accuracy is higher than 0.35 in 99% of the cases. The 95%

interval is [0.407, 0.693].

FIGURE 4: Posterior distribution of the classification accuracy. $n = 500, \lambda = 10^{-4}$.

4. DISCUSSION

Weighted Bayesian Bootstrap (WBB) provides a computationally attractive so-

lu-
tion
to
scal-
able
Bayesian
in-
fer-
ence
(**??**)
whilst
ac-
count-
ing
for
pa-
ram-
e-
ter
un-
cer-

tainty by drawing samples from a weighted posterior distribution. WBB can also be used in con- junc- tion with prox- i- mal meth- ods (**?**, **?**) to pro- vide spar- sity in high di- mensional sta- tis- tica prob- lems. With a sim- i- lar ease of com- pu- ta- tion, WBB pro- vides an alternative to ABC methods (**?**) and Variational Bayes (VB) methods. A fruitful area for fu- ture re- search is the com- par- i- son of ap- prox- i- mate Bayesian com- pu- ta- tion with simulated Bayesian Bootstrap inference.

ACKNOWLEDGEMENTS

Place all acknowledgements here. In your ini-

tial and re- vised sub- mis- sion, en- sure that any ac- knowl- edge- ments are anony- mous; in- clude the

full acknowledgements only after your paper has been accepted. Granting agencies should not

be ab- bre- vi- ated, and do not in- clude grant num- bers. We are grate- ful for your as- sistance with our publication process.

APPENDIX

APPENDIXSTOCHASTIC GRADIENT DESCENT (SGD)

Stochastic gra-

di-
ent
de-
scent
(SGD)
method
or
its
vari-
a-
tion
is
typ-
i-
cally
used
to
find
the
deep
learn-

ing model weights by minimizing the penalized loss function, $\sum_{i=1}^{n} w_i l_i(y_i; \theta) + \lambda w_p \phi(\theta)$. The method minimizes the function by taking a negative step along an estimate g^k of the gradi-

ent

$\nabla\left[\sum_{i=1}^{n} w_i l_i(y_i;\theta^k) + \lambda w_p \phi(\theta^k)\right]$

at

it-

er-

a-

tion

k.

The

ap-

prox-

i-

mate

gra-

di-

ent

is

es-

ti-

mated

by

cal-

cu-

lat-

ing

$$g^k = \frac{n}{b_k} \sum_{i \in E_k} w_i \nabla l_i(y_i; \theta^k) + \lambda w_p \frac{n}{b_k} \nabla \phi(\theta^k)$$

Where

$E_k \subset$

$\{1, \ldots, n\}$

and

$b_k =$

$|E_k|$

is

the

num-

ber

of

el-

e-

ments

in E_k. When $b_k > 1$ the algorithm is called batch SGD and simply SGD otherwise. A usual strat- egy to choose sub- set E is to go cycli- cally and pick con- sec- u- tive el- e-ments of $\{1, \ldots, T\}$, $E_{k+1} = [E_k \mod n] + 1$. The approximated direction g^k is calculated using a chain rule (aka back-propagation) for deep learning. It is an unbiased estimator.

Thus, at each it- er- a- tion, the SGD up- dates the so- lu- tion

$$\theta^{k+1} = \theta^k - t_k g^k$$

For deep learn- ing

ap-
pli-
ca-
tions
the
step
size
t_k
(a.k.a
learn-
ing
rate)
is
usu-
ally
kept
con-
stant
or
some
sim-
ple

step size reduction strategy is used, $t_k = a\exp(-kt)$. Appropriate learning rates or the hyper-

pa-

ram-

e-

ters

of

re-

duc-

tion

sched-

ule

are

usu-

ally

found

em-

pir-

i-

cally

from

nu-

merical experiments and observations of the loss function progression.

22520??

Received 9

Au-

gust

2018

Accepted 8

Septem-

ber

2018