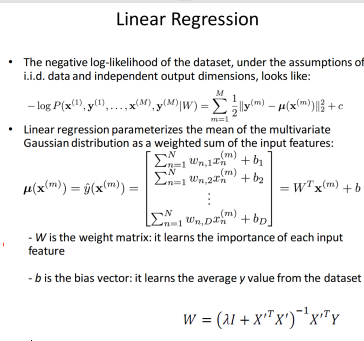
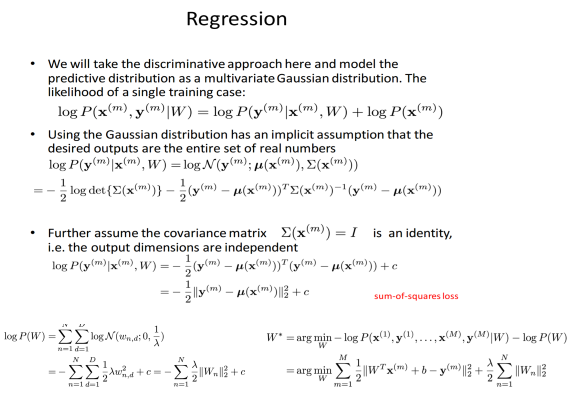


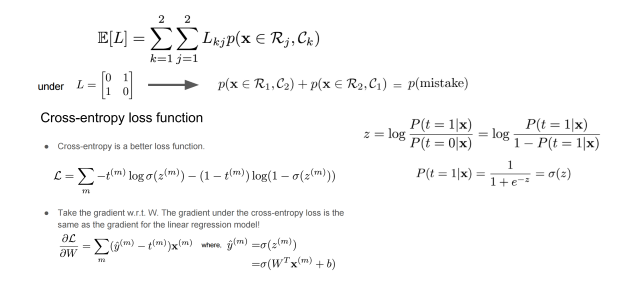
Stochastic gradient descent 
SGD is very easy to implement compared to other 
methods, but the step sizes need to be tuned to 
different problems, whereas batch learning 
typically "just works" 
Tip 1: divide the log-likelihood estimate by the size 
of your mini-batches. This makes the learning rate 
invariant to mini-batch size 
Tip 2: subsample without replacement so that you 
visit each point on each pass through the dataset 
(this is known as an epoch) 

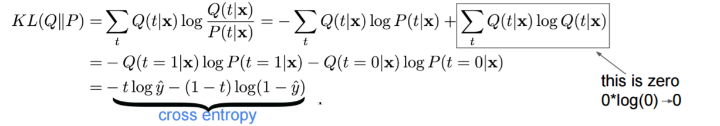


Maximum-Likelihood Regularization 
If the data are assumed to be independently and identically 
distributed (i.i.d assumption), the likelihood function takes the 
form: 
p(tlx, w, 13) 11 M (tnly(xn, w) '3 ) 
It is often convenient to maximize the log of the likelihood function: 
Inp(tlx, w, B) w) — tn)2 + — Inß 
— — In(27r). 
2 
[3E(w) 
• Maximizing log-likelihood with respect to w (under the assumption 
of a Normal noise) is equivalent to minimizing the sum-of-squared 
error function. 
• Determine WAIL by maximizing log-likelihood. Then maximizing 
1 
1 
w.r.t. ß: 
WML) — tn)2. 
[BAIL — N 

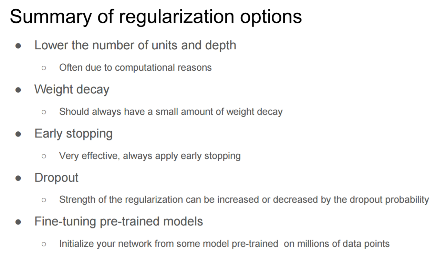


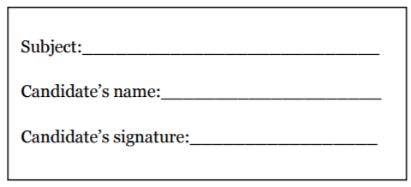
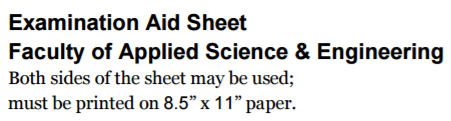
KL(QlP) 
ΚΙ αη aswnmetric 11istance tundion: K L(Qllp) K L(PIIQ) 
Q(tlx) 
KL(Qlll') —Σ Q(tlx) 
— klx) 
P(tlx) 
Μ ΤΡΕ 





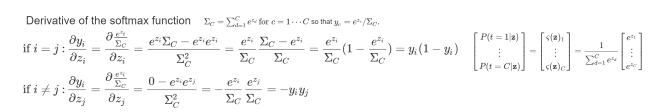
aAW(t - 1) 
vv(t) 
DC(t) 
Classical momentum 
.w(t+l) 
aAW(t - 1) 
Nesterov's 
momentum 
DC(t') 





• The hidden activation of the Jth hidden unit the second 
hidden la er is the weighted sum of the first hidden layer: 
o We can use vector notation to express the hidden vector: 
h(2) 
10 
ZH2 
x 

_ lenMe 
30 
30 Oe 
x) 



SZ

