CS 159—HW #4

Due 6 May 2021

Instructions

- The homework is scored out of 20 points.
- Complete the homework in groups of two to three, clearly listing the group members at the top of the solution. Submit only one solution per group on Gradescope.
- Either LaTeX or handwritten solutions are fine—just make sure the solution is legible. In either case turn in a pdf of your solution.
- Looking things up either in books or online is encouraged.

1 Mirror descent

Mirror descent is a framework for designing optimisation algorithms. It starts with the identity:

$$\mathcal{L}(W+\Delta W) = \underbrace{\mathcal{L}(W) + \nabla_W \mathcal{L}(W)^T \Delta(W)}_{\text{first order Taylor expansion}} + \underbrace{\mathcal{L}(W+\Delta W) - \mathcal{L}(W) - \nabla_W \mathcal{L}(W)^T \Delta W}_{\text{term that needs to be added to make this an identity}}. \tag{1}$$

where $\mathcal{L}(W) : \mathbb{R}^n \to \mathbb{R}$ is the loss function that we would like to minimise.

Since the second group of terms on the righthand side is complicated and we do not have a tractable expression for them, mirror descent models them using another function $h : \mathbb{R}^n \to \mathbb{R}$ which we are free to choose (and can choose to have nice properties). This gives us the following *model*:

$$\mathcal{L}(W+\Delta W) \overset{\text{(model)}}{=} \underbrace{\mathcal{L}(W) + \nabla_W \mathcal{L}(W)^T \Delta(W)}_{\text{first order Taylor expansion}} + \underbrace{h(W+\Delta W) - h(W) - \nabla_W h(W)^T \Delta W}_{\text{model of the second group of terms via } h}. \tag{2}$$

1.1 The mirror world (3 points)

Show that a ΔW which minimises the model in Equation 2 satisfies:

$$\nabla h(W+\Delta W) = \nabla h(W) - \nabla_W \mathcal{L}(W). \tag{3}$$

Conclude that mirror descent may be interpreted as gradient descent with step size 1 on optimisation variables transformed by the map $\nabla h(\cdot) : \mathbb{R}^n \to \mathbb{R}^n$.

Note: this gives rise to the "mirror" terminology, which is just evocative language to indicate gradient descent is occurring on transformed variables. Depending on the choice of h, ∇h might introduce a lot of distortion, so you might like to imagine the optimisation variables are being viewed in a funhouse mirror.

1.2 Strongly convex h (3 points)

Argue that if h is strongly convex, then the mirror descent update can be written in terms of the original optimisation variables as:

$$W + \Delta W = \nabla h^{-1}\big(\nabla h(W) - \nabla_W \mathcal{L}(W)\big). \tag{4}$$

Hint: what's a sufficient condition on h for ∇h^{-1} to be meaningful in the way that we want?

1.3 Entropic h (3 points)

Work out the mirror descent update (Equation 4) for the special case of $h(W) = \sum_{i=1}^n W_i \log W_i$.

2 Cover's *function-counting theorem* (5 points)

Look up the following paper and read section I:

Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition, IEEE Trans. Electron. Comput., T. Cover, 1965.

a) Explain the meaning of the term *dichotomy*, as used in the paper.

b) Argue that the total number of dichotomies of N vectors is 2^N.

c) Define the term *general position* for N vectors in d dimensions.

d) Plot the number of homogeneously linearly separable dichotomies of N points in general position in Euclidean d-space for $d = 25$ as a function of N. Put N on the horizontal axis, and let N range from 1 to 100. Normalise the vertical axis by the total number of dichotomies 2^N. *Hint: take care with how Cover defines the binomial coefficient.*

e) Explain the relevance (in broad terms) of Theorem 1 to the number of training points needed to learn a linear classifier under VC-style generalisation theory.

f) **Extra credit.** Do you expect the number of homogeneously linearly separable dichotomies of N points in Euclidean d-space to be larger or smaller if the points are *not* in general position? Start by thinking about when $d = 2$. Can you construct a general argument for arbitrary N and d?

3 Change of measure inequality (6 points)

In this question, we will derive an inequality that plays a key role in PAC-Bayesian generalisation theory.

a) Consider a positive function $\varphi : \mathbb{R}^n \to \mathbb{R}_+$, and two probability densities Q and P over $\mathbb{R}^n$ such that the support of Q satisfies $\text{supp}(Q) \subset \text{supp}(P)$. Show that:

$$\mathbb{E}_{x\sim P}[\varphi(x)] \geq \mathbb{E}_{x\sim Q}\left[\varphi(x)\frac{P(x)}{Q(x)}\right]. \tag{5}$$

b) For an arbitrary function $h : \mathbb{R}^n \to \mathbb{R}$ and an arbitrary $\beta > 0$, show that:

$$\mathbb{E}_{x\sim Q}[h(x)] \leq \frac{\text{KL}(Q||P) + \ln \mathbb{E}_{x\sim P}[\mathrm{e}^{\beta h(x)}]}{\beta}. \tag{6}$$

Hint: set $\varphi(x) = \mathrm{e}^{\beta h(x)}$ in part a) and apply Jensen's inequality.

In PAC-Bayes, we will think of P as defining a prior measure on $\mathbb{R}^n$ and Q as defining a posterior measure. We will use Inequality 6 to relate expectations under the posterior to expectations under the prior.