Machine Learning – Brett Bernstein

Week 1 Lecture: Concept Check Exercises

Starred problems are optional.

Statistical Learning Theory

1. Suppose $\mathcal{A} = \mathcal{Y} = \mathbb{R}$ and $\mathcal{X}$ is some other set. Furthermore, assume $P_{\mathcal{X}\times\mathcal{Y}}$ is a discrete joint distribution. Compute a Bayes decision function when the loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ is given by
$$\ell(a, y) = \mathbf{1}(a \neq y),$$
the $0-1$ loss.

   Solution. The Bayes decision function f^* satisfies
$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg\min_f P(f(X) \neq Y),$$
where $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$. Let
$$f_1(x) = \arg\max_y P(Y = y \mid X = x),$$
the maximum a posteriori estimate of Y. If there is a tie, we choose any of the maximizers. If f_2 is another decision function we have
$$\begin{aligned} P(f_1(X) \neq Y) &= \textstyle\sum_x P(f_1(x) \neq Y | X = x)P(X = x) \\ &= \textstyle\sum_x (1 - P(f_1(x) = Y | X = x))P(X = x) \\ &\leq \textstyle\sum_x (1 - P(f_2(x) = Y | X = x))P(X = x) \quad \text{(Defn of } f_1\text{)} \\ &= \textstyle\sum_x P(f_2(x) \neq Y | X = x)P(X = x) \\ &= P(f_2(X) \neq Y). \end{aligned}$$
   Thus $f^* = f_1$.

2. ($\star$) Suppose $\mathcal{A} = \mathcal{Y} = \mathbb{R}$, $\mathcal{X}$ is some other set, and $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ is given by $\ell(a, y) = (a - y)^2$, the square error loss. What is the Bayes risk and how does it compare with the variance of Y?

   Solution. From Homework 1 we know that the Bayes decision function is given by $f^*(x) = \mathbb{E}[Y|X = x]$. Thus the Bayes risk is given by
$$\mathbb{E}[(f^*(X) - Y)^2] = \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] = \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y|X] - Y)^2|X]] = \mathbb{E}[\mathrm{Var}(Y|X)],$$
   where we applied the law of iterated expectations. The law of total variance states that
$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathbb{E}(Y|X)].$$

This proves the Bayes risk satisfies

$$\mathbb{E}[\mathrm{Var}(Y|X)] = \mathrm{Var}(Y) - \mathrm{Var}[\mathbb{E}(Y|X)] \leq \mathrm{Var}(Y).$$

Recall from Homework 1 that $\mathrm{Var}(Y)$ is the Bayes risk when we estimate Y without any input X. This shows that using X in our estimation reduces the Bayes risk, and that the improvement is measured by $\mathrm{Var}[\mathbb{E}(Y|X)]$. As a sanity check, note that if X, Y are independent then $\mathbb{E}(Y|X) = \mathbb{E}(Y)$ so $\mathrm{Var}[\mathbb{E}(Y|X)] = 0$. If $X = Y$ then $\mathbb{E}(Y|X) = Y$ and $\mathrm{Var}[\mathbb{E}(Y|X)] = \mathrm{Var}(Y)$.

The prominent role of variance in our analysis above is due to the fact that we are using the square loss.

3. Let $\mathcal{X} = \{1, \ldots, 10\}$, let $\mathcal{Y} = \{1, \ldots, 10\}$, and let $A = \mathcal{Y}$. Suppose the data generating distribution, P, has marginal $X \sim \mathrm{Unif}\{1, \ldots, 10\}$ and conditional distribution $Y|X = x \sim \mathrm{Unif}\{1, \ldots, x\}$. For each loss function below give a Bayes decision function.

   (a) $\ell(a, y) = (a - y)^2$,
   (b) $\ell(a, y) = |a - y|$,
   (c) $\ell(a, y) = \mathbf{1}(a \neq y)$.

   Solution.

   (a) From Homework 1 we know that $f^*(x) = \mathbb{E}[Y|X = x] = (x + 1)/2$.
   (b) From Homework 1, we know that $f^*(x)$ is the conditional median of Y given $X = x$. If x is odd, then $f^*(x) = (x + 1)/2$. If x is even, then we can choose any value in the interval
   $$\left[\left\lfloor \frac{x+1}{2} \right\rfloor, \left\lceil \frac{x+1}{2} \right\rceil\right].$$
   (c) From question 1 above, we know that $f^*(x) = \arg\max_y P(Y = y|X = x)$. Thus we can choose any integer between 1 and x, inclusive, for $f^*(x)$.

4. Show that the empirical risk is an unbiased and consistent estimator of the Bayes risk. You may assume the Bayes risk is finite.

   Solution. We assume a given loss function ℓ and an i.i.d. sample $(x_1, y_i), \ldots, (x_n, y_n)$. To show it is unbiased, note that

$$\begin{aligned}
\mathbb{E}[\hat{R}_n(f)] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(f(x_i), y_i)] \qquad \text{(Linearity of } \mathbb{E}\text{)} \\
&= \mathbb{E}[\ell(f(x_1), y_1)] \qquad \text{(i.i.d.)} \\
&= R(f).
\end{aligned}$$

For consistency, we must show that as $n \to \infty$ we have $\hat{R}_n(f) \to R(f)$ with probability 1. Letting $z_i = \ell(f(x_i), y_i)$, we see that the z_i are i.i.d. with finite mean. Thus consistency follows by applying the strong law of large numbers.

5. Let $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = \mathcal{A} = \mathbb{R}$. Suppose you receive the (x, y) data points $(0, 5)$, $(.2, 3)$, $(.37, 4.2)$, $(.9, 3)$, $(1, 5)$. Throughout assume we are using the $0 - 1$ loss.

   (a) Suppose we restrict our decision functions to the hypothesis space $\mathcal{F}_1$ of constant functions. Give a decision function that minimizes the empirical risk over $\mathcal{F}_1$ and the corresponding empirical risk. Is the empirical risk minimizing function unique?

   (b) Suppose we restrict our decision functions to the hypothesis space $\mathcal{F}_2$ of piecewise-constant functions with at most 1 discontinuity. Give a decision function that minimizes the empirical risk over $\mathcal{F}_2$ and the corresponding empirical risk. Is the empirical risk minimizing function unique?

   Solution.

   (a) We can let $\hat{f}(x) = 5$ or $\hat{f}(x) = 3$ and obtain the minimal empirical risk of $3/5$. Thus the empirical risk minimizer is not unique.

   (b) One solution is to let $\hat{f}(x) = 5$ for $x \in [0, .1]$ and $\hat{f}(x) = 3$ for $x \in (.1, 1]$ giving an empirical risk of $2/5$. There are uncountably many empirical risk minimizers, so again we do not have uniqueness.

6. ($\star$) Let $\mathcal{X} = [-10, 10]$, $\mathcal{Y} = \mathcal{A} = \mathbb{R}$ and suppose the data generating distribution has marginal distribution $X \sim \text{Unif}[-10, 10]$ and conditional distribution $Y|X = x \sim \mathcal{N}(a + bx, 1)$ for some fixed $a, b \in \mathbb{R}$. Suppose you are also given the following data points: $(0, 1)$, $(0, 2)$, $(1, 3)$, $(2.5, 3.1)$, $(-4, -2.1)$.

   (a) Assuming the $0 - 1$ loss, what is the Bayes risk?

   (b) Assuming the square error loss $\ell(a, y) = (a - y)^2$, what is the Bayes risk?

   (c) Using the full hypothesis space of all (measurable) functions, what is the minimum achievable empirical risk for the square error loss.

   (d) Using the hypothesis space of all affine functions (i.e., of the form $f(x) = cx + d$ for some $c, d \in \mathbb{R}$), what is the minimum achievable empirical risk for the square error loss.

   (e) Using the hypothesis space of all quadratic functions (i.e., of the form $f(x) = cx^2 + dx + e$ for some $c, d, e \in \mathbb{R}$), what is the minimum achievable empirical risk for the square error loss.

   Solution.

(a) For any decision function f the risk is given by

$$\mathbb{E}[\mathbf{1}(f(X) \neq Y)] = P(f(X) \neq Y) = 1 - P(f(X) = Y) = 1.$$

To see this note that

$$P(f(X) = Y) = \frac{1}{20\sqrt{2\pi}} \int_{-10}^{10} \int_{-\infty}^{\infty} \mathbf{1}(f(x) = y) e^{-(y-a-bx)^2/2} \, dy \, dx = \frac{1}{20\sqrt{2\pi}} \int_{-10}^{10} 0 \, dx = 0.$$

Thus every decision function is a Bayes decision function, and the Bayes risk is 1.

(b) By problem 2 above we know the Bayes risk is given by

$$\mathbb{E}[\mathrm{Var}(Y|X)] = \mathbb{E}[1] = 1,$$

since $\mathrm{Var}(Y|X = x) = 1$.

(c) We choose $\hat{f}$ such that

$$\hat{f}(0) = 1.5, \hat{f}(1) = 3, \hat{f}(2.5) = 3.1, \hat{f}(-4) = 2.1,$$

and $\hat{f}(x) = 0$ otherwise. Then we achieve the minimum empirical risk of $1/10$.

(d) Letting

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2.5 \\ 1 & -4 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 3.1 \\ -2.1 \end{pmatrix}$$

we obtain (using a computer)

$$\hat{w} = \begin{pmatrix} \hat{d} \\ \hat{c} \end{pmatrix} = (A^T A)^{-1} A^T y = \begin{pmatrix} 1.4856 \\ 0.8556 \end{pmatrix}.$$

This gives

$$\hat{R}_5(\hat{f}) = \frac{1}{5} \|A\hat{w} - y\|_2^2 = 0.2473.$$

[Aside: In general, to solve systems like the one above on a computer you shouldn't actually invert the matrix $A^T A$, but use something like w=A\y in Matlab which performs a QR factorization of A.]

(e) Letting

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2.5 & 6.25 \\ 1 & -4 & 16 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 3.1 \\ -2.1 \end{pmatrix}$$

we obtain (using a computer)

$$\hat{w} = \begin{pmatrix} \hat{e} \\ \hat{d} \\ \hat{c} \end{pmatrix} = (A^T A)^{-1} A^T y = \begin{pmatrix} 1.7175 \\ 0.7545 \\ -0.0521 \end{pmatrix}.$$

This gives

$$\hat{R}_5(\hat{f}) = \frac{1}{5}\|A\hat{w} - y\|_2^2 = 0.1928.$$

Stochastic Gradient Descent

1. When performing mini-batch gradient descent, we often randomly choose the mini-batch from the full training set without replacement. Show that the resulting mini-batch gradient is an unbiased estimate of the gradient of the full training set. Here we assume each decision function f_w in our hypothesis space is determined by a parameter vector $w \in \mathbb{R}^d$.

   Solution. Let $(x_{m_1}, y_{m_1}), \ldots, (x_{m_n}, y_{m_n})$ be our mini-batch selected uniformly without replacement from the full training set $(x_1, y_1), \ldots, (x_n, y_n)$.

$$\begin{aligned}
\mathbb{E}\left[\nabla_w \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_{m_i}, y_{m_i}))\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\nabla_w \ell(f_w(x_{m_i}), y_{m_i})\right] && \text{(Linearity of } \nabla, \mathbb{E}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\nabla_w \ell(f_w(x_{m_1}), y_{m_1})\right] && \text{(Marginals are the same)} \\
&= \mathbb{E}\left[\nabla_w \ell(f_w(x_{m_1}), y_{m_1})\right] \\
&= \sum_{i=1}^{N} \frac{1}{N} \nabla_w \ell(f_w(x_i), y_i) \\
&= \nabla_w \frac{1}{N}\sum_{i=1}^{N} \ell(f_w(x_i), y_i) && \text{(Linearity of } \nabla).
\end{aligned}$$

2. You want to estimate the average age of the people visiting your website. Over a fixed week we will receive a total of N visitors (which we will call our full population). Suppose the population mean μ is unknown but the variance σ^2 is known. Since we don't want to bother every visitor, we will ask a small sample what their ages are. How many visitors must we randomly sample so that our estimator $\hat{\mu}$ has variance at most $\epsilon > 0$?

   Solution. Let $x_1, \ldots, x_n$ denote our randomly sampled ages, and let $\hat{x}$ denote the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i$. Then

$$\mathrm{Var}(\hat{x}) = \frac{\sigma^2}{n}.$$

Thus we require $n \geq \sigma^2/\epsilon$. Note that this doesn't depend on N, the full population size.

3. ($\star$) Suppose you have been successfully running mini-batch gradient descent with a full training set size of 10^5 and a mini-batch size of 100. After receiving more data your full training set size increases to 10^9. Give a heuristic argument as to why the mini-batch size need not increase even though we have 10000 times more data.

   Solution. Throughout we assume our gradient lies in $\mathbb{R}^d$. Consider the empirical distribution on the full training set (i.e., each sample is chosen with probability $1/N$ where N is the full training set size). Assume this distribution has mean vector $\mu \in \mathbb{R}^d$ (the full-batch gradient) and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. By the central limit theorem the mini-batch gradient will be approximately normally distributed with mean μ and covariance $\frac{1}{n}\Sigma$, where n is the mini-batch size. As N grows the entries of Σ need not grow, and thus n need not grow. In fact, as N grows, the empirical mean and covariance matrix will converge to their true values. More precisely, the mean of the empirical distribution will converge to $\mathbb{E}\nabla\ell(f(X), Y)$ and the covariance will converge to

   $$\mathbb{E}[(\nabla\ell(f(X), Y))(\nabla\ell(f(X), Y))^T] - \mathbb{E}[\nabla\ell(f(X), Y)]\mathbb{E}[\nabla\ell(f(X), Y)]^T$$

   where $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$.

   The important takeaway here is that the size of the mini-batch is dependent on the speed of computation, and on the characteristics of the distribution of the gradients (such as the moments), and thus may vary independently of the size of the full training set.

Week 1 Lab: Concept Check Exercises

Starred problems are optional.

Multivariable Calculus Exercises

1. If $f'(x; u) < 0$ show that $f(x + hu) < f(x)$ for sufficiently small $h > 0$.

   Solution. The directional derivative is given by

   $$f'(x; u) = \lim_{h\to 0} \frac{f(x + hu) - f(x)}{h} < 0.$$

   By the definition of a limit, there must be a $\delta > 0$ such that

   $$\frac{f(x + hu) - f(x)}{h} < 0$$

whenever $|h| < \delta$. If we restrict $0 < h < \delta$ then we have

$$f(x + hu) - f(x) < 0 \implies f(x + hu) < f(x)$$

as required.

2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, and assume that $\nabla f(x) \neq 0$. Prove

$$\arg\max_{\|u\|_2=1} f'(x; u) = \frac{\nabla f(x)}{\|\nabla f(x)\|_2} \quad \text{and} \quad \arg\min_{\|u\|_2=1} f'(x; u) = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}.$$

Solution. By Cauchy-Schwarz we have, for $\|u\|_2 = 1$,

$$|f'(x; u)| = |\nabla f(x)^T u| \leq \|\nabla f(x)\|_2 \|u\|_2 = \|\nabla f(x)\|_2.$$

Note that

$$\nabla f(x)^T \frac{\nabla f(x)}{\|\nabla f(x)\|_2} = \|\nabla f(x)\|_2 \quad \text{and} \quad \nabla f(x)^T \frac{-\nabla f(x)}{\|\nabla f(x)\|_2} = -\|\nabla f(x)\|_2,$$

so these achieve the maximum and minimum bounds given by Cauchy-Schwarz.

One way to understand the Cauchy-Schwarz inequality is to recall that the dot-product between two vectors $v, w \in \mathbb{R}^d$ can be written as

$$v^T w = \|v\|_2 \|w\|_2 \cos(\theta),$$

where θ is the angle between v and w. This value is maximized at $\cos(0) = 1$ and minimized at $\cos(\pi) = -1$.

3. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x, y) = x^2 + 4xy + 3y^2$. Compute the gradient $\nabla f(x, y)$.

Solution. Computing the partial derivatives gives

$$\partial_1 f(x, y) = 2x + 4y \quad \text{and} \quad \partial_2 f(x, y) = 4x + 6y.$$

Thus the gradient is given by

$$\nabla f(x, y) = \begin{pmatrix} 2x + 4y \\ 4x + 6y \end{pmatrix}.$$

4. Compute the gradient of $f : \mathbb{R}^n \to \mathbb{R}$ where $f(x) = x^T A x$ and $A \in \mathbb{R}^{n \times n}$ is any matrix.

Solution. Here we show two methods. In either case we can obtain differentiability by noticing the partial derivatives are continuous.

(a) Since

$$f(x) = x^T Ax = \sum_{i,j=1}^{n} a_{ij} x_i x_j$$

we have

$$\partial_k f(x) = \sum_{j=1}^{n} (a_{kj} + a_{jk}) x_j$$

so

$$\nabla f(x) = (A + A^T)x.$$

(b) Note that

$$\begin{aligned} f(x+tv) &= (x+tv)^T A(x+tv) \\ &= x^T Ax + tx^T Av + tv^T Ax + t^2 v^T Av \\ &= f(x) + t(x^T A + x^T A^T)v + t^2(v^T Av). \end{aligned}$$

Thus

$$f'(x;v) = \lim_{t\to 0} \frac{f(x+tv) - f(x)}{t} = \lim_{t\to 0} (x^T A + x^T A^T)v + t(v^T Av) = (x^T A + x^T A^T)v.$$

This shows

$$\nabla f(x) = (A + A^T)x.$$

5. Compute the gradient of the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = b + c^T x + x^T Ax,$$

where $b \in \mathbb{R}$, $c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$.

Solution. First consider the linear function $g(x) = c^T x$. Note that

$$g(x+tv) = c^T(x+tv) = c^T x + tc^T v \implies \nabla f(x) = c.$$

As the derivative is linear we can combine this with the previous problem to obtain

$$\nabla f(x) = c + (A + A^T)x.$$

6. Fix $s \in \mathbb{R}^n$ and consider $f(x) = (x-s)^T A(x-s)$ where $A \in \mathbb{R}^{n\times n}$. Compute the gradient of f.

Solution. We give two methods.

(a) Let $g(x) = x^T Ax$ and $h(x) = x - s$ so that $f(x) = g(h(x))$. By the vector-valued form of the chain rule we have

$$\nabla f(x) = \nabla g(h(x))^T Dh(x) = (A + A^T)(x - s),$$

where $Dh(x) = \mathbf{I}_{n\times n}$ is the Jacobian matrix of h.

(b) We have

$$(x-s)^T A(x-s) = x^T Ax - s^T(A+A^T)x + s^T As.$$

Computing the gradient gives

$$\nabla f(x) = (A+A^T)x - (A+A^T)s = (A+A^T)(x-s).$$

7. Consider the ridge regression objective function

$$f(w) = \|Aw - y\|_2^2 + \lambda \|w\|_2^2,$$

where $w \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $y \in \mathbb{R}^m$, and $\lambda \in \mathbb{R}_{\geq 0}$.

(a) Compute the gradient of f.

(b) Express f in the form $f(w) = \|Bw - z\|_2^2$ for some choice of B, z.

(c) Using either of the parts above, compute

$$\arg\min_{w \in \mathbb{R}^n} f(w).$$

Solution.

(a) We can express $f(w)$ as

$$f(w) = (Aw-y)^T(Aw-y) + \lambda w^T w = w^T A^T Aw - 2y^T Aw + y^T y + \lambda w^T w.$$

Applying our previous results gives (noting $w^T w = w^T \mathbf{I}_{n\times n} w$)

$$\nabla f(w) = 2A^T Aw - 2A^T y + 2\lambda w = 2(A^T A + \lambda \mathbf{I}_{n\times n})w - 2A^T y.$$

(b) Let

$$B = \begin{pmatrix} A \\ \sqrt{\lambda}\mathbf{I}_{n\times n} \end{pmatrix} \quad \text{and} \quad z = \begin{pmatrix} y \\ \mathbf{0}_{n\times 1} \end{pmatrix}$$

written in block-matrix form.

(c) The argmin is $w = (A^T A + \lambda \mathbf{I}_{n\times n})^{-1} A^T y$. To see why the inverse is valid, see the linear algebra questions below.

8. Compute the gradient of

$$f(\theta) = \lambda \|\theta\|_2^2 + \sum_{i=1}^n \log(1 + \exp(-y_i \theta^T x_i)),$$

where $y_i \in \mathbb{R}$ and $\theta \in \mathbb{R}^m$ and $x_i \in \mathbb{R}^m$ for $i = 1, \ldots, n$.

Solution. As the derivative is linear, we can compute the gradient of each term separately and obtain

$$\nabla f(\theta) = 2\lambda\theta - \sum_{i=1}^n \frac{\exp(-y_i\theta^T x_i)}{1 + \exp(-y_i \theta^T x_i)} y_i x_i,$$

where we used the techniques from Recitation 1 to differentiate the log terms.

Linear Algebra Exercises

1. When performing linear regression we obtain the *normal equations* $A^TAx = A^Ty$ where $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, and $y \in \mathbb{R}^m$.

   (a) If $\mathbf{rank}(A) = n$ then solve the normal equations for x.

   (b) ($\star$) What if $\mathbf{rank}(A) \neq n$?

   Solution.

   (a) We first show that $\mathbf{rank}(A^TA) = n$ to show that we can invert A^TA. By the rank-nullity theorem, we can do this by showing A^TA has trivial nullspace. Note that for any $x \in \mathbb{R}^n$ we have

   $$A^TAx = 0 \implies x^TA^TAx = 0 \implies \|Ax\|_2^2 = 0 \implies Ax = 0 \implies x = 0.$$

   This last implication follows since $\mathbf{rank}(A) = n$ so A has trivial nullspace (again by rank-nullity). This proves A^TA has a trivial nullspace, and thus A^TA is invertible. Applying the inverse we obtain

   $$x = (A^TA)^{-1}A^Ty.$$

   Since A^TA is invertible, our answer for x is unique.

   (b) We will show that the equation always has infinitely many solutions x. First note that $\mathbf{rank}(A) \neq n$ implies $\mathbf{rank}(A) < n$ since you cannot have larger rank than the number of columns. By rank-nullity, A^TA has a non-trivial nullspace, which in turn implies that if there is a solution, there must be infinitely many solutions.

   We will show that A^T and A^TA have the same column space. This will imply A^Ty is in the column space of A^TA giving the result. First note that every vector of the form A^TAx must be a linear combination of the columns of A^T, and thus lies in the column space of A^T. Above we proved that the column space of A^TA has dimension n, the same as the column space of A^T (since $\mathbf{rank}(A^T) = \mathbf{rank}(A)$). Thus A^T and A^TA have the same column spaces.

   A specific solution can be computed as $x = (A^TA)^+A^Ty$, where $(A^TA)^+$ is the *pseudoinverse* of A^TA. Of the infinitely many possible solutions x, this gives the one that minimizes $\|x\|_2$. More precisely, $x = (A^TA)^+A^Ty$ solves the optimization problem

   $$\begin{array}{ll} \text{minimize} & \|x\|_2 \\ \text{subject to} & A^TAx = A^Ty. \end{array}$$

2. Prove that $A^TA + \lambda\mathbf{I}_{n\times n}$ is invertible if $\lambda > 0$ and $A \in \mathbb{R}^{n\times n}$.

   Solution. If $(A^TA + \lambda\mathbf{I}_{n\times n})x = 0$ then

   $$0 = x^T(A^TA + \lambda\mathbf{I}_{n\times n})x = \|Ax\|_2^2 + \lambda\|x\|_2^2 \implies x = 0.$$

Thus $A^TA + \lambda \mathbf{I}_{n\times n}$ has trivial nullspace. Alternatively, we could notice that A^TA is positive semidefinite, so adding $\lambda \mathbf{I}_{n\times n}$ will give a matrix whose eigenvalues are all at least $\lambda > 0$. A square matrix is invertible iff its eigenvalues are all non-zero.

3. ($\star$) Describe the following set geometrically:

$$\left\{ v \in \mathbb{R}^2 \;\middle|\; v^T \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix} v = 4 \right\}.$$

Solution. The set is an ellipse with semi-axis lengths $2/\sqrt{3}$ and 2 rotated counter-clockwise by $\pi/4$. Letting $v = (x, y)^T$ and multiplying all terms we get

$$2x^2 + 2xy + 2y^2 = 4.$$

From precalculus we can see this is a conic section, and must be an ellipse or a hyperbola, but more work is needed to determine which one. Instead of proceeding along these lines, let's use linear algebra to give a cleaner treatment that extends to higher dimensions.

Let $A = \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}$. Since v^TAv is a number, we must have $(v^TAv)^T = v^TAv$. This gives

$$v^TA^Tv = v^TAv = \frac{1}{2}v^T(A^T + A)v = v^T \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} v.$$

Our new matrix is symmetric, and thus allows us to apply the spectral theorem to diagonalize it with an orthonormal basis of eigenvectors. In other words, by rotating our axes we can get a diagonal matrix. Either doing this by hand, or using a computer (Matlab, Mathematica, Numpy) we obtain

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = Q \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} Q^T \quad \text{where} \quad Q = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix}.$$

The set

$$\left\{ w \in \mathbb{R}^2 \;\middle|\; w^T \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} w = 4 \right\}$$

is an ellipse with semi-axis lengths $2/\sqrt{3}$ and 2 since it corresponds to the equation $3w_1^2 + w_2^2 = 4$. Since Q performs a counter-clockwise rotation by $\pi/4$ we obtain the answer. More concretely,

$$w^T \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} w = 4 \iff (Qw)^T Q \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} Q^T (Qw) = 4 \iff (Qw)^T \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} (Qw) = 4$$

so

$$\{ v \mid v^TAv = 4 \} = \left\{ Qw \;\middle|\; w^T \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} w = 4 \right\}.$$

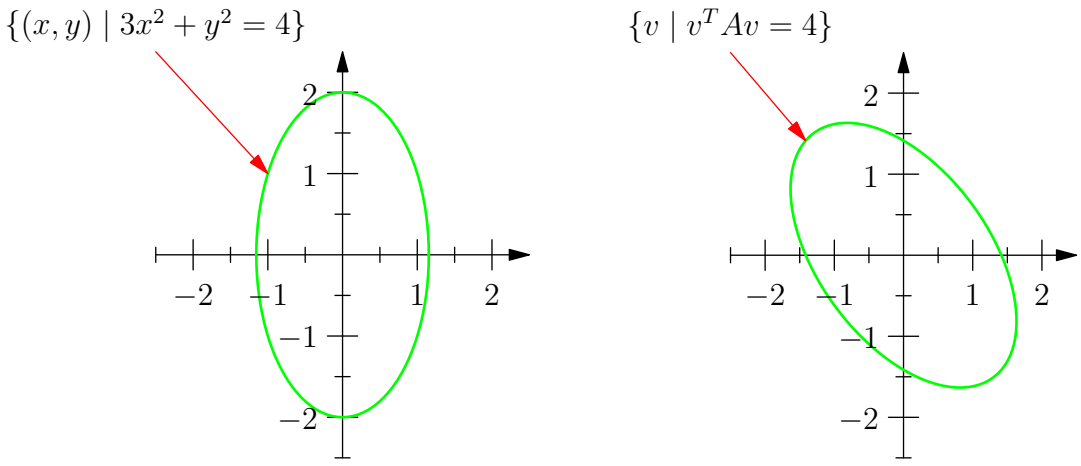

Figure 1: Rotated Ellipse

More generally, the solution to $v^T Av = c$ for $v \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ and $c > 0$ will be an ellipsoid if A is positive definite. The ith semi-axis will have length $\sqrt{c/\lambda_i}$ where λ_i is the ith eigenvalue of A.

Week 2 Pre-Lecture: Concept Check Exercises

Optimization Prerequisites for Lasso

1. Given $a \in \mathbb{R}$ we define a^+, a^- as follows:

$$a^+ = \begin{cases} a & \text{if } a \geq 0, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad a^- = \begin{cases} -a & \text{if } a < 0, \\ 0 & \text{otherwise.} \end{cases}$$

We call a^+ the *positive part* of a and a^- the *negative part* of a. Note that $a^+, a^- \geq 0$.

(a) Give an expression for a in terms of a^+, a^-.

(b) Give an expression for $|a|$ in terms of a^+, a^-.

For $x \in \mathbb{R}^d$ define $x^+ = (x_1^+, \ldots, x_d^+)$ and $x^- = (x_1^-, \ldots, x_d^-)$.

(c) Give an expression for x in terms of x^+, x^-.

(d) Give an expression for $\|x\|_1$ without using any summations or absolute values. [Hint: Use x^+, x^- and the vector $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^d$.]

Solution.

(a) $a = a^+ - a^-$

(b) $|a| = a^+ + a^-$

(c) $x = x^+ - x^-$

(d) $\|x\|_1 = \mathbf{1}^T(x^+ + x^-)$

2. Let $f : \mathbb{R} \to \mathbb{R}$ and $S \subseteq \mathbb{R}$. Consider the two optimization problems

$$\begin{array}{lll} \text{minimize}_{x\in\mathbb{R}} & |x| & \\ \text{subject to} & f(x) \in S & \end{array} \qquad \text{and} \qquad \begin{array}{ll} \text{minimize}_{a,b\in\mathbb{R}} & a+b \\ \text{subject to} & f(a-b) \in S \\ & a, b \geq 0. \end{array}$$

Solve the following questions.

(a) If $f(x) \in S$ show how to quickly compute (a, b) for the second problem with $a + b = |x|$ and $f(a - b) \in S$.

(b) If a, b satisfy $f(a-b) \in S$, show how to quickly compute an x for the first problem with $|x| \leq a + b$ and $f(x) \in S$.

(c) Assume x is a minimizer for the first problem with minimum value p_1^* and (a, b) is a minimizer for the second problem with minimum p_2^*. Using the previous two parts, conclude that $p_1^* = p_2^*$.

Solution.

(a) Let $a = x^+$ and $b = x^-$. Then $a + b = |x|$ and $a - b = x$.

(b) Let $x = a - b$ and note that $|x| = |a - b| \leq |a| + |b| = a + b$.

(c) Part a) shows $p_2^* \leq p_1^*$ by letting $\hat{a} = x^+$ and $\hat{b} = x^-$. Part b) shows $p_1^* \leq p_2^*$ by letting $\hat{x} = a - b$.

3. Let $f : \mathbb{R}^d \to \mathbb{R}$, $S \subseteq \mathbb{R}$ and consider the following optimization problem:

$$\begin{array}{ll} \text{minimize}_{x\in\mathbb{R}^d} & \|x\|_1 \\ \text{subject to} & f(x) \in S, \end{array}$$

where $\|x\|_1 = \sum_{i=1}^d |x_i|$. Give a new optimization problem with a linear objective function and the same minimum value. Show how to convert a solution to your new problem into a solution to the given problem. [Hint: Use the previous two problems.]

Solution. Consider the minimization problem

$$\begin{array}{ll} \text{minimize}_{a,b\in\mathbb{R}^d} & \mathbf{1}^T(a+b) \\ \text{subject to} & f(a-b) \in S, \\ & a_i, b_i \geq 0 \quad \text{for } i = 1, \ldots, d. \end{array}$$

Let p_1^* be the minimum for the original problem, and p_2^* the minimum for our new problem. We first show $p_1^* = p_2^*$. Suppose x is a minimizer for the original problem and let $a = x^+$ and $b = x^-$. Then by the first question $\mathbf{1}^T(a+b) = \|x\|_1$ and $a-b = x$.

This shows $p_2^* \leq p_1^*$. Next suppose (a, b) is a minimizer for our new problem, and let $x = a - b$. Then

$$\|x\|_1 = \|a - b\|_1 = \sum_{i=1}^{d} |a_i - b_i| \leq \sum_{i=1}^{d} |a_i| + |b_i| = \sum_{i=1}^{d} a_i + b_i = \mathbf{1}^T(a + b).$$

This proves $p_1^* \leq p_2^*$.

Finally, given a minimizer (a, b) for the new problem we recover a minimizer x for the original problem by letting $x = a - b$.