Machine Learning
CS229

Notes

Brendan Matthews

2025-11-24

1. Linear Regression

Let $x^{(i)} \in \mathbb{R}^n$ be the i'th feature variable of some data with m feature variables. We shall be abusing notation to write $\theta^\top x^{(i)}$ as $\theta x^{(i)}$. Assume $y^{(i)} = \theta x^{(i)} + \epsilon^{(i)} \in \mathbb{R}$ where $\epsilon^{(i)} \sim N(0, \sigma^2)$ are iid for $i = 1, ..., m$ so that $y^{(i)} \sim N\big(\theta x^{(i)}, \sigma^2\big)$. The density of this normal distribution for $y^{(i)}$ is

$$P\big(y^{(i)}|x^{(i)};\theta\big) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\theta x^{(i)}\right)^2}{2\sigma^2}}$$

and so

$$\begin{aligned}
L(\theta) &= P(y \mid x;\theta) \text{ Likelihood of } \theta \\
&= \prod_{i=1}^{m} P\big(y^{(i)}|x^{(i)};\theta\big) \text{ by our independence assumption} \\
&\implies \\
\ell(\theta) &= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\theta x^{(i)}\right)^2}{2\sigma^2}} \text{ log likelihood is monotonic so } \ell(\theta) \text{ has identical maximum to } L(\theta) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y^{(i)}-\theta x^{(i)}\right)^2}{2\sigma^2}} \\
&= m \log\Big(\sqrt{2\pi\sigma^2}\Big) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} \big(y^{(i)} - \theta x^{(i)}\big)^2.
\end{aligned}$$

Since $\sigma^2 \geq 0$[1], maximising that is equivalent to minimising

$$J : \mathbb{R}^n \to \mathbb{R}, \qquad J(\theta) = \frac{1}{2} \sum_{i=1}^{m} \big(y^{(i)} - \theta x^{(i)}\big)^2.$$

You oughta know that means you need multivariable calculus. For a function $f : \mathbb{R}^n \to \mathbb{R}$, a derivative is a gradient, a double derivative is a hessian and you might even need the chain rule.
N.B. For the express purpose of confusing people, Andrew uses avg loss sometimes. Same minimiser.

1.1. Convexity

1.1.1. Definitions and Basic Results

We need some convex analysis to understand minima properly. A function $f : \mathbb{R}^n \to \mathbb{R}_\infty$ is convex if its epigraph $\{(x,t) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq t\}$ is convex. For $\lambda \in [0,1]$, TFAE:

$$\begin{aligned}
\lambda f(x_1) + (1-\lambda) f(x_2) &\geq f(\lambda x_1 + (1-\lambda) x_2) \implies \\
\lambda t_1 + (1-\lambda) t_2 \geq \lambda f(x_1) + (1-\lambda) f(x_2) &\geq f(\lambda x_1 + (1-\lambda) x_2) \implies \\
\lambda(x_1, t_1) + (1-\lambda)(x_2, t_2) &\in \mathrm{Epi}(f) \quad \text{and} \\
\lambda(x_1, f(x_1)) + (1-\lambda)(x_2, f(x_2)) &\in \mathrm{Epi}(f) \implies \\
f(\lambda x_1 + (1-\lambda) x_2) &\leq \lambda f(x_1) + (1-\lambda) f(x_2).
\end{aligned}$$

Lemma 1.1.1.1: A twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ is convex iff it has an increasing derivative and nonnegative second derivative.

---

[1] We did not specify σ^2, so our justification for LSE is reliant strictly on the independently norm dist assumption.

Proof: Suppose f is convex. Fix $x \leq y \in \mathbb{R}$ and $0 < t < y - x$ and let $\lambda = \frac{t}{y-x}$. We have $\frac{f(y)-f(x)}{y-x} \geq \frac{f(x+t)-f(x)}{t}$ and $\frac{f(y)-f(x)}{y-x} \leq \frac{f(y)-f(y-t)}{t}$ since

$$\lambda f(y) + (1-\lambda)f(x) \geq f(x+t) = f(x+(y-x)\lambda) = f(\lambda y + (1-\lambda)x) \implies$$
$$tf(y) - tf(x) \geq (y-x)f(x+t) - (y-x)f(x).$$

and

$$\lambda f(x) + (1-\lambda)f(y) \geq f(y-t) = f(y-(y-x)\lambda) = f(\lambda x + (1-\lambda)y) \implies$$
$$tf(x) - tf(y) \geq (y-x)f(y-t) - (y-x)f(y).$$

It follows that $f'(y) \geq \frac{f(y)-f(x)}{y-x} \geq f'(x)$. So f' is increasing and f'' must be nonnegative.

Conversely, suppose f' is increasing and fix $x, z \in \mathbb{R}$ and let $y = \lambda x + (1-\lambda)z$. By MVT, $\exists (t_1, t_2) \in [x, y] \times [y, z]$ with

$$f'(t_1) = \frac{f(y) - f(x)}{y - x} \quad \text{and} \quad f'(t_2) = \frac{f(z) - f(y)}{z - y}.$$

Since $t_2 \geq t_1$, we have

$$\frac{f(z) - f(y)}{z - y} \geq \frac{f(y) - f(x)}{y - x}$$

so that f is convex. ■

Lemma 1.1.1.2: A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex iff for $x, r \in \mathbb{R}^n$, the functions $\phi : \mathbb{R} \to \mathbb{R}$ given by $\phi(t) = f(x + tr)$ are all convex.

Proof: If f is convex it's easy to see that ϕ is convex. Pick any $x, y \in \mathbb{R}^n$ and let $r = y - x$. Then by assumption the function $\phi(t) = f(x + tr)$ is convex. Then f is convex because

$$\lambda f(x) + (1-\lambda)f(y) = \lambda\phi(0) + (1-\lambda)\phi(1) \geq \phi(1-\lambda) = f(\lambda x + (1-\lambda)y).$$

■

Theorem 1.1.1.3: A twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex iff for each $x \in \mathbb{R}^n$ we have $H(f)(x) \geq 0$ where H is the Hessian matrix of partial derivatives.

Proof: Suppose that $H(f)(x) \geq 0$ for all $x \in \mathbb{R}^n$. Pick any $x, r \in \mathbb{R}^n$ and define $\phi : \mathbb{R} \to \mathbb{R}$ by $\phi(t) = f(x + tr)$. Then for $g(t) = x + tr$ we have $g'(t) = r^\top$, so

$$\begin{aligned}
\phi''(t) &= \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(x + \mathrm{tr}) \\
&= \frac{\mathrm{d}}{\mathrm{d}t}[f'(g(t))g'(t)^\top] \quad \text{multidimensional derivatives of } f, g \\
&= \frac{\mathrm{d}}{\mathrm{d}t}[g'(t)f'(g(t))^\top] \text{ because real numbers are equal to their transpose} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}[g'(t)(\nabla f)(g(t))] \\
&= r^\top \frac{\mathrm{d}}{\mathrm{d}t}[(\nabla f)(g(t))] \\
&= r^\top \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} (h_1 \circ g)(t) \\ \vdots \\ (h_n \circ g)(t) \end{pmatrix} \text{ where } h_i : \mathbb{R}^n \to \mathbb{R} \text{ is given by } h_{i(x)} = (\nabla f)_i(x) \\
&= r^\top \begin{pmatrix} h_1'(g(t))g'(t)^\top \\ \vdots \\ h_n'(g(t))g'(t)^\top \end{pmatrix} \text{ apply the derivative to the gradient by components using the chain rule} \\
&= r^\top \begin{pmatrix} h_1'(g(t)) \\ \vdots \\ h_n'(g(t)) \end{pmatrix} r \\
&= r^\top H(f)(g(t))r \\
&\geq 0.
\end{aligned}$$

Then ϕ is convex by the lemma and f is convex by the other lemma.

Now suppose that f is convex and fix $x \in \mathbb{R}^n$. Then for any $r \in \mathbb{R}^n$ the lemmas give $\phi''(0) = r^\top H(f)(x)r \geq 0$, so $H(f)(x)$ is PSD. ■

1.1.2. Unconstrained Optimisation of Convex Functions

Definition 1.1.2.1: A point $x_0 \in X$ is a local minimiser of f over X if there is a ball around x_0 such that for $y \in B_\epsilon(x_0) \cap X$ we have $f(y) \geq f(x_0)$. We say that $f(x_0)$ is a local minimum.

Theorem 1.1.2.1 (Fundamental Theorem of Convex Optimisation): Suppose f is convex with a minimum M. Then $A = \{x : f(x) \leq M\}$ is convex[2] and every local minimum is in $f(A)$. It is possible that $|A| = |\mathbb{R}|$ via

A stictly convex f has $|A| = 1$. The converse is false, $\|.\|_1$, $\|.\|_\infty$ are not strictly convex.

[2] A is a (sub)level set and $A = \{x : f(x) = M\}$. Any quasi-convex function has convex sub-level sets.

Proof: Let $f(x_0) = M$ and suppose $x_1 n = x_0$ is a local minimum in a ball $B_{\{\epsilon_0\}}$. Let $\epsilon = \min\{\epsilon_0, /\|x\|\}$. Since f is convex and $\epsilon \in (0,1)$, we have $\epsilon f(x_0) + (1-\epsilon)f(x_1) \geq f(\epsilon x_0 + (1-\epsilon)x_1)$ but $\|x\|$ and $f(\epsilon x_0 + (1-\epsilon)x_1) \geq f(x_1)$. Then $\epsilon(f(x_0) - f(x_1)) \geq 0$ and since $\epsilon > 0$ we have $f(x_0) \geq f(x_1)$ which means $f(x_0) = f(x_1) = M$. It is easy to see that if f is strictly convex, there is a contradiction, so $|A| = 1$. ■

Definition 1.1.2.2: The limit inferior of a sequence $(x_n)_{n\geq 1}$ is the largest number that eventually bounds $(x_n)_{n\geq 1}$ below.

Definition 1.1.2.3: A function $f : \mathbb{R}^n \to \mathbb{R}$ is lower semicontinuous at $x \in \mathbb{R}^n$ if for all $(x_n) \to x$ we have $f(x) \leq \liminf_{n\to\infty} f(x_n)$. A function is continuous iff it is both upper and lower semicontinuous.

Definition 1.1.2.4: A function $f : \mathbb{R}^n \to \mathbb{R}$ is coercive if $f(x_n) \to \infty$ whenever $\|x_n\| \to \infty$.

Theorem 1.1.2.2 (Extreme Value AKA Weierstrass Theorem): Suppose that C is compact and $f : C \to \mathbb{R}$ is upper/lower semicontinuous. Then f attains a maximum/minimum on C.

Corollary 1.1.2.2.1 (Existence of Solutions to Unconstrained Minimisation Problems): Let S be a nonempty and closed subset of $\mathbb{R}^n$. Suppose that $f : S \to R$ is lower semicontinuous and coercive. Then f attains a minimum on S.

Proof: If S is bounded it is compact and we're done. If not, fix $x_0 \in S$. Since f is coercive, there exists $R > 0$ such that for $\|x\| > R$ we have $f(x) \geq f(x_0)$. By Weierstrass theorem, f attains a minimum M on the compact ball $\overline{B}_{R+\|x_0\|}(x_0)$. Pick any any $x \in S$. If $x \in S \setminus \overline{B}_{R+\|x_0\|}(x_0)$ then $\|x\| + \|x_0\| \geq \|x - x_0\| > R + \|x_0\| \implies \|x\| > R$ so $f(x) \geq f(x_0) \geq M$. If not, $f(x) \geq M$ by definition. So M is the minimum of f on S. ■

Theorem 1.1.2.3 (Lipshitz Continuity of Convex Functions): Let $f : \mathbb{R}^n \to \mathbb{R}_\infty$ be a convex function and $D \subset \text{Relint}(\text{Dom}(f))$ be compact and convex. Then f is Lipshitz over D.

Everything in CS229 is continuous anyway.

1.2. The Linear Regression Cost Function is Convex

-Explainer for 2nd last line-

$$
\begin{aligned}
H(J(\theta))_{jk} &= \frac{\partial}{\partial\theta_j\theta_k}\left[\frac{1}{2}\sum_{i=1}^{m}\left(y^{(i)}-\theta x^{(i)}\right)^2\right] \\
&= \frac{\partial}{\partial\theta_j\theta_k}\left[\frac{1}{2}\sum_{i=1}^{m}\left({y^{(i)}}^2-2y^{(i)}\left(\theta x^{(i)}\right)+\left(\sum_{q=1}^{n}\theta_q x_q^{(i)}\right)^2\right)\right] \\
&= \frac{\partial}{\partial\theta_j\theta_k}\left[\frac{1}{2}\sum_{i=1}^{m}\left(\sum_{q=1}^{n}\theta_q x_q^{(i)}\right)^2\right] \\
&= \frac{\partial}{\partial\theta_j\theta_k}\left[\frac{1}{2}\sum_{i=1}^{m}\sum_{q=1}^{n}\left(\theta_q x_q^{(i)}\right)^2+\sum_{i=1}^{m}\sum_{\{r=1\}}^{m}\sum_{s=1}^{\{r-1\}}\theta_r x_r^{(i)}\theta_s x_s^{(i)}\right] \\
&= \sum_{i=1}^{m} x_j^{(i)}x_k^{(i)}. \implies H = X^\top X
\end{aligned}
$$

Since partials commute $H = H^\top$ and $z^\top H(J(\theta))z = z^\top X^\top Xz = (Xz)^\top(Xz) = (\|Xz\|_2)^2 \geq 0$, $H(J(\theta))$ is PSD and J is convex. Old mate J is continuous and looks coercive enough to me. So J has a minimum and any algorithm guaranteed to converge to a local minimum will minimise J.

1.3. Matrix Derivatives

1.3.1. Rules

Constants, sums and traces all pass through and the product rule holds.

1.3.2. Results

$$\nabla_x(x^\top b) = b$$

$$\nabla_x(x^\top Ax) = (A+A^\top)x, \qquad H(x^\top Ax) = A + A^\top \qquad \text{note: simplification when symmetric}$$

Proof: Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(x) = \langle x, Ax\rangle$ then

$$
\begin{aligned}
f(x) - f(x_0) &= \langle x, Ax\rangle - \langle x_0, Ax_0\rangle \\
&= \langle x - x_0, Ax_0\rangle + \langle x - x_0, A(x-x_0)\rangle + \langle x_0, A(x-x_0)\rangle \\
&= \langle Ax_0, x - x_0\rangle + 0.5\,\langle x - x_0, (A+A^\top)(x-x_0)\rangle + \langle A^\top x_0, x - x_0\rangle \\
&= \langle (A+A^\top)x_0, x - x_0\rangle + 0.5\,\langle x - x_0, (A+A^\top)(x-x_0)\rangle \\
\text{So } \nabla f(x) &= \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix} \\
&= \lim_{h\to 0}\frac{1}{h}\begin{pmatrix} f(x_1+h,\ldots,x_n) - f(x_1,\ldots,x_n) \\ \vdots \\ f(x_1,\ldots,x_n+h) - f(x_1,\ldots,x_n) \end{pmatrix} \\
&= \lim_{h\to 0}\frac{1}{h}(A+A^\top)hx + \frac{1}{2h}\begin{pmatrix} 2h^2A_{11} \\ \vdots \\ h^2(A_{n1}+A_{1n}) \end{pmatrix} \\
&= (A+A^\top)x
\end{aligned}
$$

■

1.4. Solving Linear Regression Explicitly

Let $X \in \mathbb{R}^{m\times n}$ be the vector with features $x^{(i)}$ as rows and let $\vec{y} \in \mathbb{R}^m$ vectorise the target variables.

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) \\
&= \frac{1}{2}\nabla_\theta \left((X\theta)^T X\theta - (X\theta)^T\vec{y} - \vec{y}^T(X\theta) + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T(X^TX)\theta - \vec{y}^T(X\theta) - \vec{y}^T(X\theta)\right) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T(X^TX)\theta - 2(X^T\vec{y})^T\theta\right) \\
&= \frac{1}{2}\left(2X^TX\theta - 2X^T\vec{y}\right) \\
&= X^TX\theta - X^T\vec{y}
\end{aligned}$$

setting to zero (see Section 5) gives the global minimiser

$$\theta = (X^\top X)^{-1}X^\top\vec{y}.$$

you need to ensure that $X^\top X$ has full rank because otherwise the kernel is not zero.

2. Logistic Regression

Suppose you have a binary classification problem. You shouldn't use linear regression. Define $h_\theta(x) = \frac{1}{1+e^{-\theta x}} \in (0,1)$. Notice that $g(x) = \frac{1}{1+e^{-x}}$ has $g'(x) = g(x)(1-g(x))$ Define h_θ as the probability of the $y^{(i)}$ taking the value of 1. Then

$$P(y|x;\theta) = h_\theta(x)^y(1-h_\theta(x))^{1-y} \implies \ell(\theta) = \sum_{i=1}^m y^{(i)}\log\left(h_\theta(x^{(i)})\right) + \left(1-y^{(i)}\right)\log\left(1-h_\theta(x^{(i)})\right)$$

which gives the loss function $J(\theta) = -\frac{1}{m}\ell(\theta)$ with $H(J(\theta)) \geq 0$. In the following, $g(X\theta) \in \mathbb{R}^m$ is treated like a row vector in numpy:

$$H_{jk} = \frac{\partial^2 J(\theta)}{\partial\theta_j\partial\theta_k} = \frac{1}{m}\sum_{i=1}^m g(\theta^T x^{(i)})[1-g(\theta^T x^{(i)})]x_j^{(i)}x_k^{(i)}$$

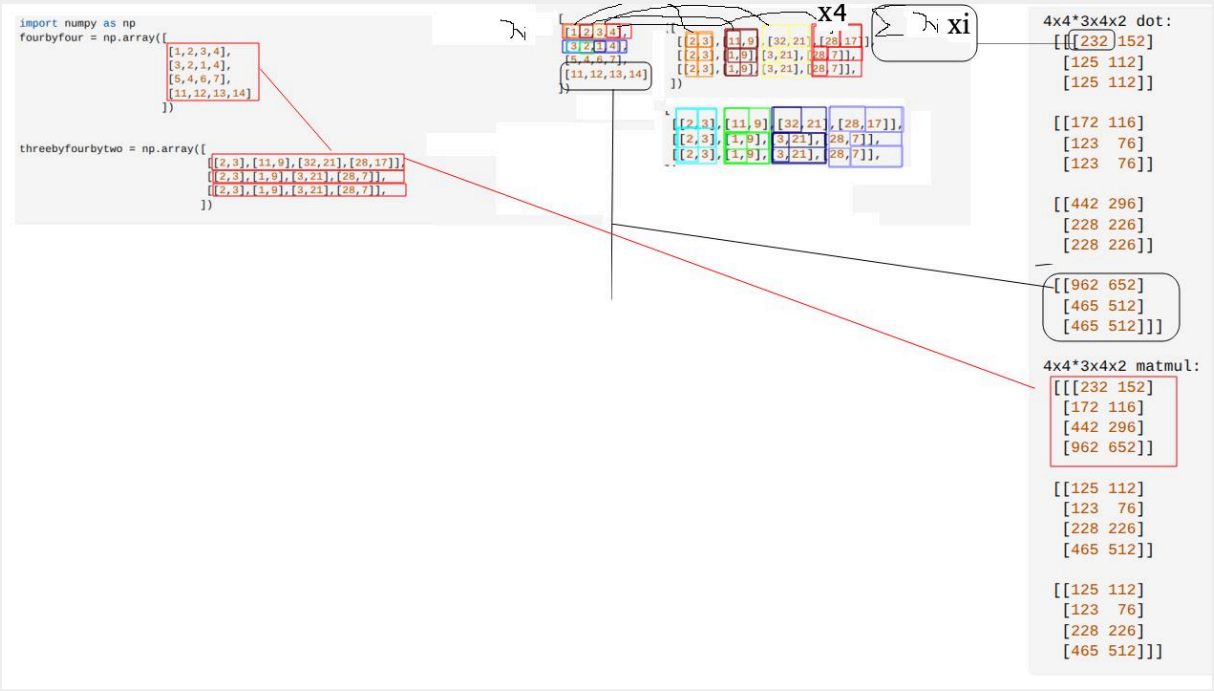
hadamard product (sort of)

$$H = \frac{1}{m}[X^T \cdot g(X\theta) \cdot (1-g(X\theta))]X$$

$$[g(\theta x^{(1)})(1-g(\theta x^{(1)})x^{(1)} \quad \dots \quad g(\theta x^{(m)})(1-g(\theta x^{(m)}))x^{(m)}]$$

$$\begin{aligned}
z^THz &= \frac{1}{m}\sum_{i=1}^m\sum_{j=1}^n\sum_{k=1}^n g(\theta^T x^{(i)})[1-g(\theta^T x^{(i)})]x_j^{(i)}x_k^{(i)}z_jz_k \\
&= \frac{1}{m}\sum_{i=1}^m g(\theta^T x^{(i)})[1-g(\theta^T x^{(i)})][(x^{(i)})^Tz]^2 \geq 0
\end{aligned}$$

N.B. The red products are those done by $*$ in numpy. The dot function is identical to numpy's matmul function for a 2D array. For multidimensional arrays, they work differently:



3. Algorithms for Solving GLM Problems

3.1. Gradient Descent

3.1.1. Batch Gradient Descent

This method is good for less than a million examples. It has slower convergence per step than Newtons method but the steps are less costly. The update rule minimising J is

$$\text{While True: } \theta := \theta - \alpha \nabla J(\theta) = \theta + \alpha \sum_{i=1}^{m} \big(y^{(i)} - h_\theta\big(x^{(i)}\big)\big) x^{(i)}$$

where $\alpha > 0$ is the learning rate. This also maximises the likelihood of θ because log is increasing.

3.1.2. Stochastic Gradient Descent

This method is good for more than a million examples. It has less reliable convergence than batch gradient descent but the steps are less costly. The update rule is

$$\text{While True: For } i = 1, \ldots, m; \quad \theta := \theta + \alpha\big(y^{(i)} - h_\theta\big(x^{(i)}\big)\big) x^{(i)}$$

Guaranteed convergence is still possible by decaying α.

3.2. Newtons Method

Suppose you wanna find the zero of a function f. Pick an $x_1 \in \mathbb{R}$. Set $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$ and you will converge on the zero. Assuming one global extremum, Newtons method applied to f' will find x such that f is minimised or maximised. In our case for $\ell(\theta)$, we maximise via the update rule

$$\theta := \theta - H^{-1}(\ell)(\theta) \nabla \ell(\theta) \text{ unless } \nabla \ell(\theta) = 0 \ldots$$

It is fairly obvious that there is no notion of ascent or descent like there is in the gradient algorithms because the zeros of $-f$ and f coincide. **We require that $H(\ell)(\theta)$ is positive/negative definite**.

4. These algorithms are well and good but why do they work?

Definition 4.1 (Directional Derivative): Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\phi(t) = f(x + td)$. We say that f has a directional derivative at $x \in \mathbb{R}^n$ in the direction $d \in \mathbb{R}^n$ if $\phi'(0) = \langle \nabla f(x), d \rangle$ exists.

Definition 4.14 (descent direction) *Let the function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be given. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a vector such that $f(\boldsymbol{x})$ is finite. Let $\boldsymbol{p} \in \mathbb{R}^n$. We say that the vector $\boldsymbol{p} \in \mathbb{R}^n$ is a descent direction with respect to f at $\boldsymbol{x}$ if*

$$\exists \delta > 0 \text{ such that } f(\boldsymbol{x} + \alpha \boldsymbol{p}) < f(\boldsymbol{x}) \text{ for every } \alpha \in (0, \delta]$$

holds. ■

4.1. Gradient Descent

4.1.1. Why Does the Algorithm give a Descent direction?

Fix $J \in (\mathbb{R}^n)^{\mathbb{R}} \cap C^1$, $\theta \in \mathbb{R}^n$ and $\epsilon > 0$. From now on we will denote $-\nabla J(\theta)$ by v. Define $\phi \in \mathbb{R}^{\mathbb{R}}$ by $\phi(t) = J(\theta + tv)$. Since $J \in C^1$, the Taylor series for ϕ about 0 gives

$$\begin{aligned} J(\theta + tv) = \phi(t) &= \phi(0) + t\phi'(0) + o(t) \\ &= J(\theta) + t \langle \nabla J(\theta), v \rangle + o(t) \\ &= J(\theta) - t\|v\|^2 + o(t) \\ &< J(\theta) \text{ for small enough } t > 0 \text{ and } v \neq 0. \end{aligned}$$

The above calculation shows that if $\nabla J(\theta) \neq 0$ then there is a small enough step in the direction of $-\nabla J(\theta)$ that decreases J, so θ is not a local minimum. The contrapositive yields Theorem 5.1.

4.1.2. What Does Steepest Descent Even Mean?

By definition, the angle φ between a descent direction p and the vector $\nabla J(\theta)$ is given by

$$\cos(\varphi) = \frac{\langle \nabla J(\theta), p \rangle}{\|\nabla J(\theta)\| \|p\|}.$$

When $p = -\alpha \nabla J(\theta)$, we have

$$\cos(\varphi) = \frac{-\alpha \langle \nabla J(\theta), \nabla J(\theta) \rangle}{\alpha \|\nabla J(\theta)\| \|\nabla J(\theta)\|} = -1,$$

so that $\langle \nabla J(\theta), p \rangle = \cos(\varphi) \|\nabla J(\theta)\| \|p\|$ is minimised. We call this the direction of steepest descent.

4.2. Newton's Method

4.2.1. Why Does the Algorithm give a Descent direction?

Fix $J \in (\mathbb{R}^n)^{\mathbb{R}} \cap C^2$, $\theta \in \mathbb{R}^n$ and $\epsilon > 0$. From now on we will denote $-H^{-1}(J(\theta)) \nabla J(\theta)$ by v. Define $\phi \in \mathbb{R}^{\mathbb{R}}$ by $\phi(t) = J(\theta + tv)$. Since $J \in C^2$, the Taylor series for ϕ about 0 gives

$$\begin{aligned} J(\theta + tv) = \phi(t) &= \phi(0) + t\phi'(0) + o(t) \\ &= J(\theta) + t\left\langle \nabla J(\theta), v \right\rangle + o(t) \\ &= J(\theta) - t\left\langle \nabla J(\theta), H^{-1}(J(\theta))\nabla J(\theta) \right\rangle + o(t) \\ &< J(\theta) \text{ for small enough } t > 0 \text{ and } \nabla J(\theta) \neq 0 \text{ since } \ H^{-1} > 0. \end{aligned}$$

We can replace $H^{-1} > 0$ with any $Q > 0$. **N.B the learning rate must be sufficiently small to guarantee descent; however this is not done in practice, see Section 4.2.5.** What if $\nabla J(\theta) = 0$ at a saddle point?

4.2.2. How do we Modify Descent Algorithms at STATIONARY Points

Fix $J \in (\mathbb{R}^n)^{\mathbb{R}} \cap C^2$, $\theta \in \mathbb{R}^n$ and $\epsilon > 0$. Assume that $H(J(\theta))$ is **not** positive semidefinite, if it is then θ typically a local minimiser, but it might not be unless H is PD. Let v be an eigenvector of $H(J(\theta))$ with eigenvalue $\lambda < 0$. Define $\phi \in \mathbb{R}^{\mathbb{R}}$ by $\phi(t) = J(\theta + tv)$. Since $J \in C^2$, the Taylor series for ϕ about 0 gives

$$\begin{aligned} J(\theta + tv) = \phi(t) &= \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(0) + o(t^2) \\ &= J(\theta) + t\left\langle \nabla J(\theta), v \right\rangle + \frac{t^2}{2}\left\langle H(J(\theta))v, v \right\rangle + o(t^2) \\ &= J(\theta) + \frac{\lambda t^2}{2}\|v\|^2 + o(t^2) \\ &< J(\theta) \text{ for small enough } t > 0. \end{aligned}$$

4.2.3. Why is Newton's Method Any Good?

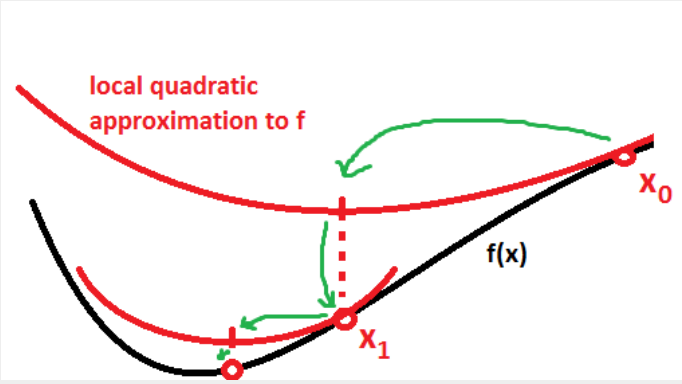
If $J \in (\mathbb{R}^n)^{\mathbb{R}} \cap C^2$ then the second order approximation of $\phi(t) = J(\theta + tp)$ gives

$$\begin{aligned} \phi(1) &\approx \phi(0) + \phi'(0) + \frac{1}{2}\phi''(0) \\ \text{'Amount of Ascent'} &= J(\theta + p) - J(\theta) \\ &\approx \varphi_x(p) = \left\langle \nabla J(\theta), p \right\rangle + \frac{1}{2}\left\langle H(J(\theta))p, p \right\rangle. \end{aligned}$$

The quadratic approximation $\phi_x(p)$ is strictly convex since $H > 0$. We have

$$\nabla \phi_x(p) = \nabla J(\theta) + H(J(\theta))p.$$

Setting this to zero yields $p = -H^{-1}(J(\theta))\nabla J(\theta)$. So the amount of descent of the quadratic approximation is maximised when $\alpha = 1$ in Newton's method.



4.2.4. In What Sense is it the Steepest Descent Direction?

Any symmetric matrix defines an inner product $\langle x, y\rangle_A = \langle x, Ay\rangle$ and a norm too if A is PD. Let $f : \left(\mathbb{R}^n, \langle ., .\rangle_{\text{generic}}\right) \to \mathbb{R}$. Fix $x \in \mathbb{R}^n$. Since $Df(x) : \left(\mathbb{R}^n, \langle ., .\rangle_{\text{generic}}\right) \to \mathbb{R}$ given by

$$Df(x)[d] = \lim_{t\to 0} \frac{f(x+td) - f(x)}{t}$$

is a continuous linear functional the Riesz representation theorem says there is a unique vector g such that

$$\langle d, g\rangle_{\text{generic}} = Df(x)[d] \qquad \forall d \in \mathbb{R}^n.$$

We call g the gradient of f at the point x. In the Hessian geometry, the gradient of f at x is $g = H^{-1}(f)(x)\nabla f(x)$ since

$$\langle d, g\rangle_H = \langle d, \nabla f(x)\rangle_{\text{2-norm}} = \lim_{t\to 0} \frac{f(x+td) - f(x)}{t}.$$

The step size in Newtons method is $\left\|H^{-1}(f)(x)\nabla f(x)\right\|$, which is the norm of the gradient in the Hessian geometry: $\|\Delta\theta\|_H = \left\|H^{-1}(f)(x)\nabla f(x)\right\|_H = \|g\|_H$.[3]. The function value changes according to $\frac{1}{2}\nabla f(x)^\top H^{-1}(f)(x)\nabla f(x) = \frac{1}{2}\left\|H^{-1}(f)(x)\nabla f(x)\right\|_H^2 = \frac{1}{2}\|g\|_H^2$ by Section 4.2.3 and so the method updates the function value according to the magnitude of the square of the gradient in the Hessian geometry, the same as gradient descent does in Euclidean geometry.

4.2.5. Convergence

Section 4.2.1

5. Why Does Setting the Gradient to Zero Work?

The gradient is zero at minimisers of C^2 functions and points where the gradient is zero are minimisers if the Hessian is locally PD or the function is convex.

Theorem 5.1 (Necessary Conditions for Minimisers of C1 & C2 Functions): Let $f \in (\mathbb{R}^n)^{\mathbb{R}} \cap C^1$. If θ is a local minimiser of f, then $\nabla f(\theta) = 0$. Furthermore, if $f \in C^2$ then for $w \in \mathbb{R}^n$ and $t > 0$

$$\begin{aligned} f(\theta + tw) - f(\theta) = \phi(t) - f(\theta) &= \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(0) - f(\theta) + o(t^2) \\ &= t\left\langle \nabla f(\theta), w\right\rangle + \frac{t^2}{2}\left\langle H(f(\theta))w, w\right\rangle + o(t^2) \\ &= \frac{t^2}{2}\left\langle H(f(\theta))w, w\right\rangle + o(t^2) \\ &\geq 0 \text{ for small enough } t \text{ because } \theta \text{ is a local minimiser.} \end{aligned}$$

So $H(f(\theta))$ is positive semidefinite. If θ is a strict local minimiser, then $H(f(\theta))$ is PD.

- f need not be convex.
- Converse is false, $f(x) = x^3$ has $f'(0) = 0$ and $f''(0) = 0$ but 0 is not a local minimiser.

Proof: For the C^1 condition, see Section 4.1.1. ∎

[3] there is a cool duality that $\|\nabla f(x)\|_{H^{-1}} = \|\Delta\theta\|_H$

Corollary 5.1.1 (Sufficient Conditions for Minimisers of C2 Functions): Let $f \in (\mathbb{R}^n)^{\mathbb{R}} \cap C^2$. If $\nabla f(\theta) = 0$ and $H(f(\theta))$ is PD then θ is a strict local minimiser. This only works for $H(f(\theta)) = \epsilon > 0$ because we can control t so that $o(t^2) < \epsilon$.

The C^1 calculation in Section 4.1.1 gives us a bit more, if we think of $v \neq 0$ as an arbitrary vector then $J(\theta + tv) = J(\theta) + t \langle \nabla J(\theta), v \rangle + o(t) < J(\theta)$ for small enough t if $\langle \nabla J(\theta), v \rangle < 0$. So v is a decent direction.

Theorem 5.2: Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable everywhere. Then f is convex iff

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \iff \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0.$$

Proof: If f is convex then for $\lambda \in (0,1)$, $\lambda(f(y) - f(x)) \geq f(x + \lambda(y-x)) - f(x)$ so $f(y) - f(x) \geq \frac{f(x+\lambda(y-x)) - f(x)}{\lambda} \geq \lim_{\lambda \to 0^+} \frac{f(x+\lambda(y-x)) - f(x)}{\lambda} = \langle \nabla f(x), y - x \rangle$
Conversely...

$$\begin{aligned} \langle \nabla f(y), y - x \rangle &\geq f(y) - f(x) \\ -\langle \nabla f(x), y - x \rangle &\geq f(x) - f(y) \\ \implies & \\ \langle \nabla f(y) - \nabla f(x), y - x \rangle &\geq 0. \end{aligned}$$

Let $\phi(t) = f(x + t(y - x))$. Fix $t_2 > t_1$. Then

$$\begin{aligned} \phi'(t_2) - \phi'(t_1) &= \langle \nabla f(x + t_2(y-x)), y - x \rangle - \langle \nabla f(x + t_1(y-x)), y - x \rangle \\ &= \frac{1}{t_2 - t_1} \langle \nabla f(a) - \nabla f(b), a - b \rangle \geq 0. \end{aligned}$$

Then by Lemma 1.1.1.1, ϕ is convex and so f is too by Lemma 1.1.1.2. ■

Corollary 5.2.1 (Sufficient Conditions for Minimisers of C1 Convex Functions): Let $J : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and convex. If $\nabla J(\theta) = 0$ then $\theta = \text{Argmin}(J)$ since $J(\varphi) \geq J(\theta) + \langle \nabla J(\theta), \varphi - \theta \rangle = J(\theta)$.

6. Generalised Linear Models

A GLM consists of

- a distribution from the exponential family,
- a linear predictor

$$\eta(\theta) = \begin{pmatrix} \theta_1^\top \\ \vdots \\ \theta_k^\top \end{pmatrix} x \in \mathbb{R}^k \text{ with } \eta_i = \theta_i^\top x \text{ for } i = 1, ..., k \text{ and } \theta = (\theta_1, ..., \theta_k)^\top \in \mathbb{R}^{nk} \text{ with } \theta_i \in \mathbb{R}^n.$$

- a link function g such that $E(T(y)|x) = \nabla A(\eta)$ provided $A \neq 0$.
- It is true that $\text{Var}(T(y); \eta) = H(A)(\eta)$ provided $A \neq 0$.

The canonical link function is $g = \nabla A$ where A is the log-partition function. The density of an exponential family distribution is

$$P(y;\eta) = b(y)e^{\langle \eta, T(y)\rangle - A(\eta)}$$

which has nothing to do with the price of fish. What we care about is

$$\ell(\theta) = \log \prod P\big(y^{(i)}|x^{(i)};\theta\big) = \sum_{i=1}^{m} b\big(y^{(i)}\big) + \left\langle \begin{pmatrix} \theta_1^\top x^{(i)} \\ \vdots \\ \theta_k^\top x^{(i)} \end{pmatrix}, T\big(y^{(i)}\big) \right\rangle - A \circ \begin{pmatrix} \theta_1^\top x^{(i)} \\ \vdots \\ \theta_k^\top x^{(i)} \end{pmatrix}.$$

The Bernoulli distribution has

$$\eta = \log\left(\frac{\phi}{1-\phi}\right) \qquad T(y) = y, \qquad b(y) = 1 \qquad A(\eta) = \log(1+e^\eta).$$

While training a model with the ASSUMPTION $P(y=1|x;\theta) = \phi$, we completely ignore $\eta = \log\big(\frac{\phi}{1-\phi}\big)$ this is used to determine ϕ for new examples once θ is optimised, which then allows classification.

The Poisson distribution $\frac{\lambda^y e^{-\lambda}}{y!} = \frac{1}{y!}e^{y\log(\lambda)-\lambda}$ has

$$\eta = \log(\lambda), \qquad T(y) = y, \qquad b(y) = \frac{1}{y!}, \qquad A(\eta) = e^\eta.$$

The distribution $N(\mu, 1)$ has

$$\eta = \mu, \qquad T(y) = y, \qquad b(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}, \qquad A(\eta) = \frac{\eta^2}{2}.$$

The log-likelihood function is confusingly written as

$$\ell(\theta) = \sum_{i=1}^{m} \log\big(b\big(y^{(i)}\big)\big) + \big(\eta^{(i)}\big)^\top T\big(y^{(i)}\big) - A\big(\eta^{(i)}\big).$$

We first find the transpose of the Jacobian of $\eta^{(i)} : \mathbb{R}^{nk} \to \mathbb{R}^k$ given by $\eta_j^{(i)}(\theta) = \theta_j^\top x^{(i)}$:

$$\begin{aligned}
\Big(\eta^{(i)\prime}(\theta)\Big)^\top &= J^\top\big(\eta^{(i)}\big)(\theta) \\
&= \begin{pmatrix} \frac{\partial\eta_1}{\partial\theta_{11}} & \cdots & \frac{\partial\eta_1}{\partial\theta_{1k}} & & \frac{\partial\eta_1}{\partial\theta_{n1}} & \cdots & \frac{\partial\eta_1}{\partial\theta_{nk}} \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \frac{\partial\eta_k}{\partial\theta_{11}} & \cdots & \frac{\partial\eta_k}{\partial\theta_{1k}} & & \frac{\partial\eta_k}{\partial\theta_{nk}} & \cdots & \frac{\partial\eta_k}{\partial\theta_{nk}} \end{pmatrix}^\top \\
&= \begin{pmatrix} \nabla_{\theta_1}\eta_1 & & \\ & \ddots & \\ & & \nabla_{\theta_k}\eta_k \end{pmatrix}_{nk\times k} \\
&= \begin{pmatrix} x^{(i)} & & \\ & \ddots & \\ & & x^{(i)} \end{pmatrix}_{nk\times k}
\end{aligned}$$

The grad chain rule says that for $f : \mathbb{R}^k \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}^k$ we have $\nabla(f\circ g) = J(g)^\top(\nabla f)\circ g$.

$$
\begin{aligned}
\nabla \ell(\theta) &= \sum_{i=1}^{m} \nabla_\theta \delta\big(\eta^{(i)}(\theta)\big), \qquad \text{with } \delta : \mathbb{R}^k \to \mathbb{R} \text{ given by } \;\; \delta(x) = x^\top T\big(y^{(i)}\big) - A(x) \\
&= \sum_{i=1}^{m} J^\top\big(\eta^{(i)}\big)(\theta) \nabla \delta\big(\eta^{(i)}\big) \\
&= \sum_{i=1}^{m} \begin{pmatrix} x^{(i)} & & \\ & \ddots & \\ & & x^{(i)} \end{pmatrix}_{nk \times k} \big(T\big(y^{(i)}\big) - \nabla A\big(\eta^{(i)}\big)\big) \iff \sum_{i=1}^{m} \big(T\big(y^{(i)}\big) - A'\big(\eta^{(i)}\big)\big) x^{(i)} \text{ when } k = 1. \\
&= \sum_{i=1}^{m} \big(T\big(y^{(i)}\big) - \nabla A\big(\eta^{(i)}\big)\big) \otimes x^{(i)} \quad \text{(Kronecker Product)}
\end{aligned}
$$

The Hessian chain rule says that for $f : \mathbb{R}^k \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}^k$ we have

$$
H(f \circ g)(x) = J(g)(x)^\top H(f)(g(x)) J(g)(x) + \sum_{i=1}^{k} (\nabla f)_i (g(x)) H(g_i)(x).
$$

So

$$
\begin{aligned}
H(\ell)(\theta) &= \sum_{i=1}^{m} H\big(\delta \circ \eta^{(i)}\big)(\theta) \\
&= \sum_{i=1}^{m} J^\top\big(\eta^{(i)}\big)(\theta) H(\delta)\big(\eta^{(i)}\big) J\big(\eta^{(i)}\big)(\theta) + \sum_{j=1}^{k} (\nabla \delta)_j \big(\eta^{(i)}\big) H\Big(\eta_j^{(i)}\Big)(\theta) \\
&= \sum_{i=1}^{m} \begin{pmatrix} x^{(i)} & & \\ & \ddots & \\ & & x^{(i)} \end{pmatrix}_{nk \times k} H(\delta)\big(\eta^{(i)}\big) \begin{pmatrix} x^{(i)} & & \\ & \ddots & \\ & & x^{(i)} \end{pmatrix}^\top_{k \times nk} + \sum_{j=1}^{k} \big(T\big(y^{(i)}\big) - \nabla A\big(\eta^{(i)}\big)\big)_j \times 0_{nk \times nk} \\
&= -\sum_{i=1}^{m} \begin{pmatrix} x^{(i)} & & \\ & \ddots & \\ & & x^{(i)} \end{pmatrix}_{nk \times k} H(A)\big(\eta^{(i)}\big) \begin{pmatrix} x^{(i)} & & \\ & \ddots & \\ & & x^{(i)} \end{pmatrix}^\top_{k \times nk} \\
&= -\sum_{i=1}^{m} \begin{pmatrix} x_1 h_{11} & \dots & x_1 h_{1k} \\ & \vdots & \\ x_n h_{11} & \dots & x_n h_{1k} \\ & \vdots & \\ & \vdots & \\ x_1 h_{k1} & \dots & x_1 h_{kk} \\ & \vdots & \\ x_n h_{k1} & \dots & x_n h_{kk} \end{pmatrix} \begin{pmatrix} x_1 & \dots & x_n & & & & & & \\ & & & x_1 & \dots & x_n & & & \\ & & & & & & \ddots & & \\ & & & & & & & x_1 & \dots & x_n \end{pmatrix} \\
&= -\sum_{i=1}^{m} \begin{pmatrix} x_1^2 h_{11} & \dots & x_1 x_n h_{11} & \dots & x_1^2 h_{1k} & \dots & x_1 x_n h_{1k} \\ & & & \vdots & & & \\ x_n x_1 h_{11} & \dots & x_n^2 h_{11} & \dots & x_1 x_n h_{1k} & \dots & x_n^2 h_{1k} \\ & & & \vdots & & & \\ & & & \vdots & & & \\ x_1^2 h_{k1} & \dots & x_1 x_n h_{k1} & \dots & x_1^2 h_{kk} & \dots & x_1 x_n h_{kk} \\ & & & \vdots & & & \\ x_n x_1 h_{k1} & \dots & x_n^2 h_{k1} & \dots & x_1 x_n h_{kk} & \dots & x_n^2 h_{kk} \end{pmatrix}
\end{aligned}
$$

$$= -\sum_{i=1}^{m} H(A)\big(\eta^{(i)}\big) \otimes x^{(i)} {x^{(i)}}^{\top} \in \mathbb{R}^{nk \times nk} \simeq \mathbb{R}^{k \times k} \otimes \mathbb{R}^{n \times n}$$

$$\iff -\sum_{i=1}^{m} \mathrm{Var}\big(T\big(y^{(i)}\big); \eta^{(i)}\big) x^{(i)} {x^{(i)}}^{\top} \leq 0 \text{ when } k = 1 \text{ since } -\sum_{i=1}^{m} \mathrm{Var}\big(T\big(y^{(i)}\big); \eta^{(i)}\big) w^{\top} x^{(i)} \big(w^{\top} x^{(i)}\big)^{\top} \leq 0.$$

So 1D GLM's are concave but higher dimensional GLM's need not be.

6.1. 2D Example: Categorical Distribution

Suppose $y^{(i)} \in \{1, ..., k\}$ and $P\big(y^{(i)} = i | x^{(i)}; \theta\big) = \phi_i$ where $\sum \phi_i = 1$ and $\eta_i = \theta_i^{\top} x^{(i)}$. The dimensionality of the distribution is $k - 1$ since $\phi_k = 1 - \sum_{i=1}^{k-1} \phi_i$. For $T(y) = e_y \in \mathbb{R}^{k-1}$ we have

$$P\big(y^{(i)} = k\big) = \phi_k^{1 - \sum_{i=1}^{k-1} (y=i)} \prod_{i=1}^{k-1} \phi_i^{(y^{(i)}=i)} = \exp\left(\left\langle \log \begin{pmatrix} \frac{\phi_1}{1 - \sum \phi_i} \\ \vdots \\ \frac{\phi_{k-1}}{1 - \sum \phi_i} \end{pmatrix}, T\big(y^{(i)}\big) \right\rangle + \log\left(1 - \sum_{i=1}^{k-1} \phi_i\right) \right)$$

so $\eta_j = \log\Big(\frac{\phi_j}{1 - \sum_{i=1}^{k-1} \phi_i}\Big)$, $T\big(y^{(i)}\big) = e_{y^{(i)}}$ and $A(\eta) = \log\Big(1 + \sum_{j=1}^{k-1} e^{\eta_j}\Big)$. The log-likelihood function is

$$\ell(\theta) = \sum_{i=1}^{m} \left[\log\left(1 + \sum_{j=1}^{k-1} e^{\eta_j^{(i)}}\right) + \sum_{j=1}^{k} \big(y^{(i)} = j\big) \eta_j^{(i)} \right]$$

6.2. Properties of Canonical Exponential Family

The canonical k-dim exponential family has probability density

$$p(x; \theta) = h(x) e^{\sum_{i=1}^{k} [\theta_i T_i(x)] - A(\theta)}$$

We assume in this section that the distribution is canonicial. The joint distribution of a bunch of iid exp. dist. random variables is exp. dist. with sufficient statistic $\sum_i T_i(X_i)$. The Hessian of the log partition function is equal to the covariance matrix of $T(X)$, i.e. $H(A)_{ij} = \mathrm{Cov}\big(T(X)_i, T(X)_j\big)$ which is positive definite. The Hessian of the log likelihood of m iid samples is then $-mH(A)$ which is concave. So a GLM is concave in the natural parameters.

7. Averages of independent random variables concentrate around their expectation

Some course objectives for students in machine learning include: (1) Predict which kinds of existing machine learning algorithms will be most suitable for which sorts of tasks, based on formal properties and experimental results. (2) Evaluate and analyze existing learning algorithms.

Definition 7.1: For discrete random variables $X, Y : \Omega \to \mathbb{R}$ we define

$$P(X = Y) = \sum_{\lambda \in X(\Omega)} P(X^{-1}\{\lambda\} \cap Y^{-1}\{\lambda\}).$$

A CDF of a random variable $X : \Omega \to \mathbb{R}$ is a function $F : \mathbb{R} \to [0, 1]$,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} F'(y)\,\mathrm{d}y \quad \text{or} \quad \sum_{\lambda \leq x} P(X = \lambda) \text{ where } \quad P(X = \lambda) = P(X^{-1}\{\lambda\})$$

is a right continuous[4] function such that $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

2.11 Definition. *A random variable X is* **continuous** *if there exists a function f_X such that $f_X(x) \geq 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and for every $a \leq b$,*

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx. \tag{2.2}$$

The function f_X is called the **probability density function** (PDF)**.** *We have that*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

and $f_X(x) = F'_X(x)$ at all points x at which F_X is differentiable.

If the sample space Ω is countable, X is discrete with PMF given by the probabilities of preimages as above. This is NOT true of a PDF, which can be unbounded at some points. We say that $X \overset{d}{=} Y$ if $F_X = F_Y \not\Rightarrow X = Y$.

The Inverse CDF is given by $F^{-1}(q) = \inf\{x \in \mathbb{R} : F(x) > q\}$.

[4] nondecreasing of course

Definition/Theorem 7.2: We define $\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)p_X(x)\,\mathrm{d}x$ and $\mathrm{Var}(g(X)) = \mathbb{E}[(gX - \mathbb{E}(gX))^2]$ and cov is defined easily from that. Correlation is cov divided by the product of the variances. From the above, we get the law of iterated expectation:

$$\begin{aligned}\mathbb{E}(\mathbb{E}(Y|X)) &= \int_{\mathbb{R}} \mathbb{E}(Y|X=x)p_X(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} yp_{Y|X}(y|x)p_X(x)\,\mathrm{d}x\,\mathrm{d}y \text{ by Fubini } (\mathbb{R} \text{ is } \sigma\text{-finite}) \\ &= \int_{\mathbb{R}} y\left(\int_{\mathbb{R}} p_{X,Y}(x,y)\,\mathrm{d}x\right)\mathrm{d}y \text{ by defn of } p_{Y|X}(y|x) \\ &= \mathbb{E}(Y) \text{ by the marginal distribution.}\end{aligned}$$

and

$$\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}(Y|X)) + \mathrm{Var}(\mathbb{E}(Y|X))$$

since

$$\mathbb{E}(\mathrm{Var}(Y|X)) = \mathbb{E}(\mathbb{E}(Y^2|X) - \mathbb{E}(Y|X)^2) = \mathbb{E}(Y^2) - \mathbb{E}(\mathbb{E}(Y|X)^2)$$

$$\mathrm{Var}(\mathbb{E}(Y|X)) = \mathbb{E}(\mathbb{E}(Y|X)^2) - \mathbb{E}(Y)^2.$$

The red terms are random variables, not fixed numbers. In particular,

$$\mathbb{E}(Y|X) : \Sigma \to \mathbb{R}, \qquad \mathbb{E}(Y|X)(\omega) = \mathbb{E}(Y|X = X(\omega))$$

Definition 7.3: The MGF of an RV is a negative double sided Laplace transform:

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_{\mathbb{R}} e^{tx}p_X(x)\,\mathrm{d}x$$

We have $M_X^{(n)}(0) = \mathbb{E}(X^n)$ since

$$\frac{d^n}{dt^n}\mathbb{E}(e^{tX}) \overset{\text{dom. conv.}}{=} \mathbb{E}\left(\frac{d^n}{dt^n}e^{tX}\right) = \mathbb{E}(X^n e^{tX}).$$

The characteristic function is the Fourier transform $\mathbb{E}(e^{itX})$.

7.1. CDFs of Transformations

For $Y = g(X)$,

$$F_Y(y) = P(Y \le y) = P(x \text{ given } g(x) \le y) = \int_{A(y)} p_X(x)\,\mathrm{d}x.$$

Take $Y = g(X) = \log(X)$ and $p_X(x) = e^{-x},\ x > 0$. Then

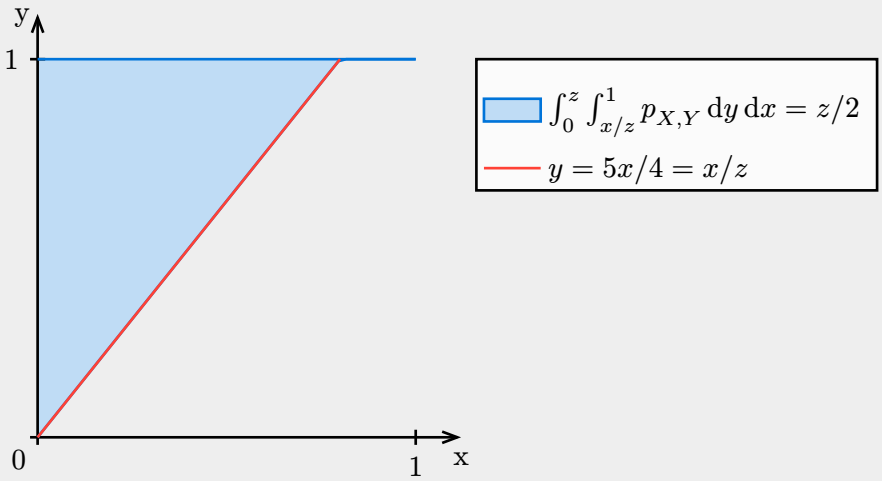
$$A(y) = \{x : \log(x) \le y\} = \{x : x \le e^y\}$$

$$F_Y(y) = \int_0^{e^y} e^{-x}\,\mathrm{d}x = 1 - e^{-e^y}, \qquad p_Y(y) = e^y e^{-e^y}, \qquad y \in \mathbb{R}.$$

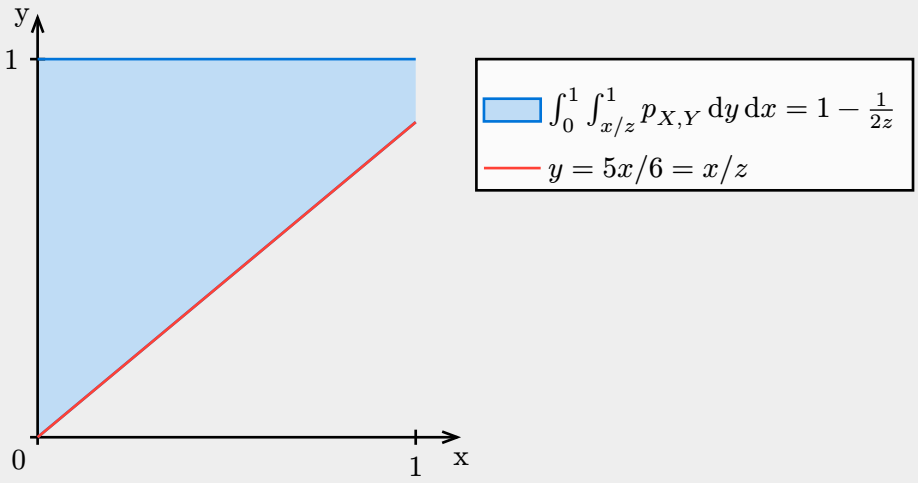
Similar notions make sense if $g : \mathbb{R}^n \to \mathbb{R}^{m \leq n}$, but the region A is harder to integrate over. Take independent $X, Y \sim \mathrm{Unif}(0,1)$ and $Z = X/Y$ so that

$$p_{X,Y}(x,y) = \{0 < x < 1\}\{0 < y < 1\} \quad \text{and} \quad F_Z(z) = F_X(Z^{-1}) = \int\int_{\{x \leq zy\}} p_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

The density weighted volume under the region with this particular density is just its area. If $z \leq 0$, the region is empty so the area is zero. Fix z between 0 and 1, say $z = 0.8$. The region in 3D space we are integrating under is $d : x \geq 0 \wedge x \leq 1 \wedge y \geq 0 \wedge 1 \geq y \wedge 0.8y \geq x$. The region reaches as far as z in the x direction.



For $z \geq 1$, say $z = 1.2$, the region extends to one in the x direction:



Another thing we can do is use the vector to vector transform when $g : \mathbb{R}^n \to \mathbb{R}^n$

Vector to vector [edit]

Suppose $\mathbf{x}$ is an n-dimensional random variable with joint density f. If $\boldsymbol{y} = G(\boldsymbol{x})$, where G is a bijective, differentiable function, then $\boldsymbol{y}$ has density $p_{\boldsymbol{Y}}$:

$$p_Y(\mathbf{y}) = f\Big(G^{-1}(\mathbf{y})\Big) \left| \det \left[\frac{dG^{-1}(\mathbf{z})}{d\mathbf{z}} \bigg|_{\mathbf{z}=\mathbf{y}} \right] \right|$$

with the differential regarded as the Jacobian of the inverse of $G(\cdot)$, evaluated at $\boldsymbol{y}$.[7]

to get an expression for $p_{Z,Y}$ and then marginalise the joint PDF to get p_Z and F_Z. Consider the transform $G(X,Y) = (X/Y, Y)$. The formula above yields

$$p_G(g_1, g_2) = p_{X,Y}(g_1 g_2, g_2) \left| \det \begin{pmatrix} g_2 & g_1 \\ 0 & 1 \end{pmatrix} \right| = p_{X,Y}(g_1 g_2, g_2)|g_2| = p_{Z,Y}(g_1, g_2),$$

so

$$\begin{aligned} f_Z(z) &= \int_{\mathbb{R}} p_{Z,Y}(z,y)\,\mathrm{d}y \\ &= \int_{\mathbb{R}} p_{X,Y}(zy,y)|y|\,\mathrm{d}y \\ &= \int_{\mathbb{R}} \{0<zy<1\}\{0<y<1\}|y|\,\mathrm{d}y \\ &= \{z>0\}\int_0^{\min(1,1/z)} y\,\mathrm{d}y \\ &= \begin{cases} 0, & z \le 0 \\ \frac{1}{2}, & 1 > z > 0 \\ \frac{1}{2z^2}, & z \ge 1 \end{cases} \end{aligned}$$

and

$$F_Z(z) = \begin{cases} 0, & z \le 0 \\ z/2, & 1 > z > 0 \\ 1 - \frac{1}{2z}, & z \ge 1 \end{cases}$$

7.2. Distributions

7.2.1. Normal

Ye old

$$\frac{1}{(2\pi)^{\frac{1}{d}}|\Sigma|} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$$

As expected $A \times N(\mu, \Sigma) + b = N(A\mu + b, A\Sigma A^\top)$. As not so expected $M(t) = e^{\mu^\top t + \frac{t^\top \Sigma t}{2}}$.

Theorem 7.2.1.1: Suppose that $X_1, ..., X_n \sim N(\mu, \sigma)$[5]
- $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
- S^2 and $\overline{X}$ are independent.

These are not obvious.

7.2.2. Chi Squared

The sum of squares of p standard normal distributions $Y = \left(Y_1, ..., Y_p\right)$ has Chi-squared dist parameter p. If the normal distributions are not standard but have unit variance $Z \sim N\left(\nu, I_p\right)$ then the sum of squares has a non-central Chi-squared distribution $Z^\top Z \sim \chi^2_p(\nu^\top \nu)$. If $Y \sim N(\mu, \Sigma)$, the covariance matrix is symmetric PSD and has an eigen decomposition that allows square roots, which leads to $Y^\top \Sigma^{-1} Y \sim \chi^2_p(\mu^\top \Sigma^{-1} \mu)$ via $Z = \Sigma^{-\frac{1}{2}} Y$, $Z \sim N\left(\Sigma^{-\frac{1}{2}}\mu, 1\right)$ by symmetry of $\Sigma^{-\frac{1}{2}}$.

7.2.3. Multinomial

$p(x) = \binom{n}{x_1, ..., x_k} p_1^{x_1}, ..., p_k^{x_k}$, where $x_1, ..., x_k$ are the numbers of cards drawn from each of the k categories after n draws.

7.3. Tail Behaviour of RV's

Theorem 7.3.1 (Markov Inequality): Suppose $X \geq 0$ and fix $t > 0$. Then

$$\mathbb{E}(X) = \int_0^\infty x p(x)\,\mathrm{d}x \underset{\text{since } p \geq 0}{\geq} t \int_t^\infty p(x)\,\mathrm{d}x = t\mathbb{P}(X \geq t).$$

Corollary 7.3.1.1 (Chebyshev's Inequality): Viewing $(X - \mathbb{E}(X))^2$ as a random variable,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) = \mathbb{P}((X - \mathbb{E}(X))^2 \geq t^2) \leq \frac{\sigma^2}{t^2} \text{ where } \ t^2 = \lambda > 0,$$

provided that $\mathbb{E}\left(X^2\right)$ is finite. Similarly,

$$\mathbb{P}\left((X - \mathbb{E}(X))^k \geq t^k\right) \leq \frac{\mathbb{E}\left((X - \mathbb{E}(X))^k\right)}{t^k}$$

given that $\mathbb{E}\left(X^k\right)$ is finite.

Example 7.3.1: Suppose that $X_1, ..., X_n \sim N(\mu, \sigma)$. Then

$$\mathbb{P}\left(\left|\overline{X} - \mu\right| \geq t\right) \leq \frac{\mathrm{Var}\left(\overline{X}\right) + \mathbb{E}^2\left(\overline{X}\right) - \mu^2}{t^2} = \frac{\sigma^2}{nt^2}$$

and also $\mathbb{P}\left(\left|\overline{X} - \mu\right| \geq \frac{t\sigma}{\sqrt{n}}\right) = \frac{1}{t^2}$.

[5] This notation apparently implies independence

These work well on nasty sums of random variables too. The ones that follow don't.

7.3.1. We can do a lot better with the additional assumption that RV's are Independent. We need to know that $M_X(t) = e^{t\mu + \frac{1}{2}(t\sigma)^2}$ for normal distributions.

Theorem 7.3.1.1 (Chernoff's Bound): Suppose that M_X is bounded in a neighbourhood of the origin $[-b, b]$. Then for any $t \in [0, b]$

$$\mathbb{P}(X - \mu \geq u) = \mathbb{P}\big(e^{t(X-\mu)} \geq e^{tu}\big) \leq \frac{\mathbb{E}\big(e^{t(X-\mu)}\big)}{e^{tu}} = \frac{M_X(t)}{e^{t(u+\mu)}}$$

and since the infimum is the greatest lower bound,

$$\mathbb{P}(X - \mu \geq u) \leq \inf_{0 \leq t \leq b} e^{-t(\mu+u)} M_X(t).$$

Example 7.3.1.1: Suppose that $X \sim N(\mu, \sigma)$. Then the MGF is finite for $t \geq 0$ so by calculus

$$\mathbb{P}(X - \mu \geq u) \leq \inf_{0 \leq t < \infty} e^{-t(\mu+u)} e^{t\mu + \frac{1}{2}(t\sigma)^2} = \inf_{0 \leq t < \infty} e^{-tu + \frac{1}{2}\sigma^2 t^2} = e^{-\frac{u^2}{2\sigma^2}}$$

and

$$\mathbb{P}(-X + \mu \geq u) \leq \inf_{0 \leq t < \infty} e^{-t(-\mu+u)} e^{-t\mu + \frac{1}{2}(t\sigma)^2} = \inf_{0 \leq t < \infty} e^{-tu + \frac{1}{2}\sigma^2 t^2} = e^{-\frac{u^2}{2\sigma^2}}$$

Then since $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$,

$$\mathbb{P}(|X - \mu| \geq u) \leq 2e^{-\frac{u^2}{2\sigma^2}}$$

and in particular since $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ we have

$$\mathbb{P}\Big(\big|\overline{X} - \mu\big| \geq t\Big) \leq 2e^{-\frac{nt^2}{2\sigma^2}}$$

and also $\mathbb{P}\left(\big|\overline{X} - \mu\big| \geq \frac{t\sigma}{\sqrt{n}}\right) \leq 2e^{-t^2/2}$. These bounds control the difference between a sample mean and mean of iid normal RV's much better than Chebyshev.

Definition 7.3.1.1 (Sub-Gaussian RV's): The same bounds hold for any RV with $\mathbb{E}\big(e^{t(X-\mu)}\big) \leq e^{\frac{(t\sigma)^2}{2}}$. Rademacher RV's are an example.

7.4. Dimensionality Reduction