Proofs Related to Lagrange Multipliers

I Introduction

1. We base ourselves on the survey of proofs given in [@c1].

II Dabbling I

2. If we think about the problem of optimizing linear functions constrained by linear functions, we see that it does not make much sense, since the intersection of a plane in $\mathbb{R}^3$ (the surface on which to optimize) and a line (a constraint) is a line with usually unbounded value in height.

3. We note that we can to begin with restrict ourselves with the cases where the constraining sets, once intersected with the surface on which to optimize, reduce the dimension of it, but not to zero, (e.g optimize a parabola for y conditioned to the intersection with a line, which leaves two points which we can simply check to choose the best)

4. We then have to think about the nonlinear differentiable case.

5. Note that thinking of the problem in terms of the resulting constraint curve (set) first, we are trying to optimizing a function on that curve, isn't this just calculus on manifolds? Can we turn this into a cartesian coordinate setting somehow?

6. The first thought is re-parametrisation. We could get rid of the constraints. If the constraint can be turned into an explicit function, we can do substitution and this gives us an unconstrained problem. This is probably the reason for the use of the implicit theorem in some proofs. As we know this might not be possible, and as we know, the usual trick is to resort to thinking in terms of the linear approximations. The idea is that extrema have almost local properties, and this could be reflected in the linear approximations. Indeed this seems to give us a good idea. At a point on the constraint curve, we can move only along the tangent to the curve. We can check if this movement helps us extremise the function to optimize by looking at the normal vector to it, since it points in the way of the extremum (its projection being the gradient). Basically, if the projection of one on the other is not zero, we can move along the tangent while making the value larger. At the extremum we cannot do that and the projection is zero. This could be the gist of it.

7. A good faithful picture is a parabola in $\mathbb{R}^3$ intersecting a skewed plane as a constraint. We are optimizing 'on the resulting ellipse' for the value of the optimisation function. Note that if the plane is horizontal, all points are extrema. The details of this can be worked out, we shall see if this is good enough. Specifically, the number and relation between the dimensions have to be worked out, and it must be related to 'some' gradients in either $\mathbb{R}^3$ or projected on the $\mathbb{R}^2$ plane, and turning a tangent into a normal can turn the orthogonality condition into a parallelism condition.

III Refinement of $\lim$ I

8. In the handwritten note 'LagrProof[3]', we find an opportunity to once more refine our idea of limits, even more than the one in 'pointillistic'.

9. We arrive to the conclusion that $\lim_{h\to 0}$ is already too loaded to be conceptually clear.

10. We would like to think of the limit as much as we can as a function. A function that as a domain has a set, but one that has a specified cluster point. If we take some ball around the cluster point in $\mathbb{R}^n$ as one such set, clearly, the ball has multiple cluster points, so the ball is not enough to describe the point. We need both the ball and the cluster point to be specified. Given this, we see inner balls of different radii as level curves. We would like to see the limit as a function acting on the set, turning the level curves into other (level) curves. But a function will not do here, because the limit could split the specific cluster point into two distinct ones (this is a case where the limit does not converge (per example, for a step function, the abscissa a is mapped by the limit (not by the original function) to the two different step point, for this reason, there will be a level curve of nonzero size at the output side of the limit that cannot be made smaller, or in other words, the single point is exploded into multiple, and the level curves must accommodate, the infinitesimal last curve become one with finite radius holding multiple points). So the limit is though of as a relation. In fact, only when it is a function can the limit converge.

11. If we take the point of view above, and look a the case of domain being $\mathbb{R}^1$, the level curves become endpoints of closed intervals. For the level curve $|h| = 1$, we then look where $(-1, 1)$ (as an ordered set of two points, not as an interval) is mapped, etc. Another important thing to notice is that the function (and hence the limit) at -1 could be undefined, and this is yet another reason why we probably cannot rigorously think of things that way, but almost, since at some point, the function must start being defined.

12. Looking at this this way, we notice that the step function is in a certain way continuous. It is both left-continuous and right-continuous. We were brought to this thinking by our goal of thinking about extrema, whose property is exactly expressed by comparing both left and right sides of the extremum. But this means that left-continuity is a more primitive concept that continuity.

13. We start thinking about paths. If a path towards the cluster point in the domain is such that it is mapped to a path that tends to some cluster point in the co-domain, do we have a limit? Yes, for that path! For every co-level curve (which is in this case a sample point on the path), we have a level-curve that would allow us the output the drive close to the input.

Here we must first pause to clear a misconception. We were still surprisingly unsure that this was the case. Given $a > b$ in the domain, mapping to $f(a) > f(b)$ in the codomain, since all we do is sampling (sequences), we were afraid the proof would not say anything about some violating c between a, b whose $f(c)$ is such that $f(c) > f(a)$, but this can not happen. What saves us and allows us to prove with sequences while we really mean the whole continuum of the path is the exact definition of convergence of a sequence, which guarantees that for all p point less than a , it holds $f(a) > f(p)$. We never paid attention to this. This is a strong condition that allows us to use sequences although we mean continue (no holes). This is expressed in the handwritten sketch (a). So then maybe the atoms of thinking about limits should be path based. As we saw, the proof using a 'sampled' sequence on a path is a proof for any other sequence on that path. However, it is not a proof for any other path. Indeed, the definition of continuity 'elegantly' (which usually means obfuscatingly to the beginner), dictates both paths to be proven. Since if we are working on the path from the right in the step continuous function, the 'there is a delta' forces an interval, which includes the left path, because we are talking directly

about level-curves (intervals) and not paths. In a better, less elegant, way we could talk about the right path, the left path, and if they are equal, we have 'continuity', that is 'left-right' continuity, that is 'ball' continuity, that is 'absolute value' continuity, that is 'level-curve' continuity. We can see how this is exactly the issue when we pass to n-var calculus . Partial derivatives are 'left-right' limits but only for one 'line-path'. The proof of differentiability is then about proving that continuous partials do give support to 'all paths'! (e.g one step on first partial path and one on the second partial path).

14. With the help of this point of view, it seems we have the right mental refinement, since this made us paid attention to details that felt 'technical' before but not anymore.

15. Maybe just like limits can erase strictness, they can also turn functions into relations (hence once possible source of divergence).

16. What we think we uncovered here is already treated 'elegantly' in a way that is so sad that we seem not to remember it at all! Our 'paths' are continuous versions of sequences. Our note about a 'violating c between a, b ' while true, does not go far enough. Paths can be as wild as we like, oscillating, etc., more so in n-var. But this is where the forgotten theorem comes in:

$$\lim_{x \to a} f(x) = L$$

if and only if ***for all sequences*** x_n (with x_n not equal to a for all n) converging to a the sequence $f(x_n)$ converges to L. It was shown by Sierpiński in 1916 that proving the equivalence of this definition and the definition above, requires and is equivalent to a weak form of the axiom of choice. Note that defining what it means for a sequence x_n to converge to a requires the epsilon, delta method.

This means that sequences are 'sampled paths', and this makes sense since it relates to our expressivity (finite). This means that sequences can be used to find counter-examples, to prove divergence. Such a counter-example samples a path that leads to a different point (turns the function into a relation as we said, explodes the limit) than the one sought. It feels psychologically and intuitively futile to try to prove something ***for all sequences***, but this is an illusion and misconception. One can per example, reason on all sequences converging to 0, per example, by saying that the element-wise squares of all these sequences also converge to 0. Per example, taking $\sqrt(x)\cos(1/x)$, we look at $\sqrt(x_n)\cos(1/x_n)$ for all sequences that converge to zero, and we can say that this converges to zero, for all sequences.

17. More importantly than the above very late misconception, we notice that as we said, limits are unavoidable as expressions of most quantities (e.g derivative) since irrationals will appear as possible values, and hence, even though technically we often end up looking at limits, this is the wrong 'focus', it is not about the limits, it is about the structure of each specific problem to be studied, and that is the fun in each problem. There is no point searching for a method, because this is impossible. The limit is not important, each problem will carry its own questions and structure, even if they will seem to be passing through limits, and reducing to finding or testing them.

18. In fact, this refinement allows us to see limits as a general technique to approach defining 'things with properties' that are not possible to define directly, but as limits. As soon as one understands this refinement, one sees it everywhere:

1. The vec-calc definition of derivative [VLDU]

2. A version of the definition of the Lebesgue integral [AMP p.80]
3. A validation of our independetly found general instance of it in [cont_short]: "x is a ***limit*** of y: x is join-recoverable from y after its removal from the latter."

19. In our case, we saw in 'LagrProof[2], LagrProof[3](a)' how the structure of the two things to compare is the essence of the problem, the thing to focus us, the think that makes the problem itself and not any other problem.

IV Proof Sketch of 'Extrema are Critical'

20. ***Proof Sketch ('LagrProof[1-3]').*** Given a function f with x^* as an extremum within a domain D . Notice the formal similarity between:

$$E(x,h) = f(x^*+h) - f(x^*), \tag{1}$$

$$f'(x) = \lim_{h\to 0} \frac{f(x^*+h) - f(x^*)}{h}, \tag{2}$$

21. and that $E(x,h)$ does not change sign within D. In a multivariate setting, having chosen any direction with h parametrizing walking along it, $E(x,h)/h$ changes sign at $h = 0$. We use this to restrict the partial derivatives of f along any direction (to zero). We are working with the assumption that the f is differentiable at x^* . For any direction, the limits from both sides for h must agree:

$$\lim_{h\to 0^+} \cdots = \lim_{h\to 0^-} \ldots .$$

22. Clearly, any sequence of secants on one side having the same sign means that the limits must have that sign or be equal to zero —This is obvious because if the limit has sign opposite to the sequence of secants, any ball of radius half the limit around the limit will not contain any element of the sequence—. Since the limits from both sides must agree while at the same time cannot have different signs, the limits must be zero. All partial derivatives must be zero, x^* must be a critical point. □

V 'Lagrange Multipliers', Proof Sketch I

23. ***Proof Sketch ('LagrProof[4-5]').*** Consider the special case of a single objective function $z = o(x_i)$ and its surface $\mathcal{O}$ constrained by a planar surface $c(x_i) = 0$ and its surface $\mathcal{C}$.

24. The admissible set (intersection curve) of extrema is $\mathcal{A} = \mathcal{O} \cap \mathcal{C}$. In the cases where one can obtain an explicit solution for $\mathcal{A}$, per example by substitution, we are back to the unconstrained case, but this is not a general solution. What we will try to do is find relations on the differential level that allow us to solve the problem despite $\mathcal{A}$ being implicitly given, having no explicit form.

25. Let us think about $\mathcal{A}$ as an unconstrained problem. The usual condition can be see as that the tangent to $\mathcal{A}$ at an extremum p^* be orthogonal to the 'objective direction' z.

26. Looking at the problem from the point of view of $\mathcal{A}$, we find no difference between the roles of the objective surface and the constraint surface. We need to extremise with respect to z, which is independent of both, and the intersection favors none of them. The only difference being that extremisation requires an explicitly expressed objective function.

27. Let us try to related the condition at p^* that $\vec{a} \perp \vec{z}$ to conditions on some differential quantities pertaining to our given surfaces. By $\vec{a}$ we mean a tangent vector to $\mathcal{A}$ at a point and by $\vec{z}$ the unit direction along which we extremise.

28. First of all, we always have that at any point of $\mathcal{O}, \mathcal{C}$:

$$\vec{a} \perp \overset{\perp}{\mathcal{O}}, \tag{3}$$

$$\vec{a} \perp \overset{\perp}{\mathcal{C}}, \tag{4}$$

where $\overset{\perp}{\mathcal{X}}$ is the surface normal vector to $\mathcal{X}$ at the point in question

29. At p^*, $\vec{a} \in \perp_{\vec{z}}$ implies that

$$\overset{\perp}{\mathcal{O}}, \overset{\perp}{\mathcal{C}} \in (\overset{\perp}{\perp_{\vec{z}}}). \tag{5}$$

Hence at an extremum, $\overset{\perp}{\mathcal{O}}$ and $\overset{\perp}{\mathcal{C}}$ reside in a hyper-plane orthogonal to z, their projections on the x_i hyperplanes are parallel, they differ only by a scalar, the Lagrange multiplier. □

30. A figure supporting the sketch is http://tex.stackexchange.com/questions/286653 http://austingwalters.com/edge-detection-in-computer-vision/

VI Proof Sketch of Multivariable 'Extrema are Critical'

31. After this, we go back to formalizing the lagrproof sketch I, whose step $\vec{a} \in \perp_{\vec{z}}$ is now formalized.

VII Tangent Plane in Terms of Jacobian (and Notation)

32. First of all, there is no mention anywhere of Tangent planes (let alone tangent spaces which is already in the abstract setting) when one is talking about partial derivatives of a function, and the vector thereof, the Jacobian. It is necessary to finally memorize that the multivariable function setting is not the surface in space setting. In the former, there is a Jacobian, and not a tangent plane —since for a function $\mathbb{R}^n \to \mathbb{R}$ the 'tangent plane to the curve' would reside in the 'geometric space' $\mathbb{R}^{n+1}$ which does not figure—, but a linear approximation L to the function. This is clear in [VLDU](p.114). It is obvious that the curve setting is a generalization of the function setting, and hence has its additional theorems and structures, such as the gradient which again in the space of its curve, not in the domain of some function, even if we can treat a function as a curve but not vice versa. The treatment of the function setting is still treated separately because indeed it is a good stepping stone to the general case, and it can use the properties of something being a function and not a relation.

Next we dig up our notation during the investigation of continuous partials, we treat our misunderstanding of what exactly is continuous, especially in the light of (13), then we pass on to expressing [VLDU](p.114) in our notation.

33. While digging up the notation, and before even finding it, we see that we closed a loop that we started during the investigation of the cotangent space and covecs while finalizing multilinear algebra with [CItLA]. We note the following necessary internalizations from p.551 where we read

If $\{x^i\}$ is a local coordinate system on a differentiable manifold X , then a (tangent) vector field $v(x)$ on X is defined as the derivative function $v = v^i(\partial/\partial x^i)$, so that $v(f) = v^i(\partial f/\partial x^i)$ for every smooth function $f : X \to \mathbb{R}$ (and where each v^i is a function of position $x \in X$, i.e., $v^i = v^i(x)$). Since every vector at $x \in X$ can in this manner be written as a linear combination of the $\partial/\partial x^i$, we see that $\{\partial/\partial x^i\}$ forms a basis for the tangent space at x .

1. We read "If $\{x^i\}$ is a local coordinate system". We finally have stopped seeing x_i as the fuzzy idea of a 'variable' and in the diff-geom setting at least as a proper (abstract) vector (in general we could see it as a formal or logical symbol). In the abstract setting, (where there is no 'ambient space' $\mathbb{R}^n$) it is a very big deal to bring a vector space and a manifold into relation. It is a big deal to speak of a set of vectors x_i as the basis about which we are going to coordinate an otherwise abstract lonesome manifold X. Let us recall that the broadest definition of a manifold is that of a ***topological space locally iso (homeomorphic) to a vector space over the specific topological field*** $\mathbb{R}$.
2. We should also see $v = v^i(\partial/\partial x^i)$ as what exactly? A function that is a vector? We again close a loop here by looking at [ItSM] which describes tangent vector to smooth manifolds in the abstract sense and provides a clearer (and longer, since it is the focus of the book) treatment of $v = v^i(\partial/\partial x^i)$ as seen this this note. Therefore, we next treat tangent vectors on smooth manifolds in the treatment VIII. The focus after coming back should be on the so-far ignored but crucial role of coordinate functions in [ItSM] (p.53) where we read

   and taking f to be the jth coodinate function $x^j : \mathbb{R}^n \to R$.

VIII Tangent Vectors on Smooth Manifolds with Lee.

34. How to write the definition on [Lee, p.3] in an internalisable way? First we gather what we need in the short [@c5], then we add our notes here...

IX Intuitions with Karabulut and with Klein

35. A comment on Karabulut's infinitesimal argument can be found here: http://math.stackexchange.com/questions/674

X Proof with Brezhneva

36. bla

Bibliography

Brezhneva, Olga, Alexey A Tret'yakov, and Stephen E Wright. 2012. "A Short Elementary Proof of the Lagrange Multiplier Theorem." ***Optimization Letters*** 6 (8). Springer: 1597–1601.

Karabulut, Hasan. 2007. “The Physical Meaning of Lagrange Multipliers.” ***ArXiv Preprint ArXiv:0705.0609***. Citeseer.

Klein, Dan. 2004. “Lagrange Multipliers Without Permanent Scarring.” ***University of California at Berkeley, Computer Science Division***.

Lee, John M. 2003. “Smooth Manifolds.” In ***Introduction to Smooth Manifolds***, 1–29. Springer.

Nohra, Jad. “Smooth Manifolds in Short.”