Gradients and Hessians in Hilbert spaces

Jordan Bell
jordan.bell@gmail.com
Department of Mathematics, University of Toronto

June 19, 2015

1 Gradients

Let $(X, \langle \cdot, \cdot \rangle)$ be a real Hilbert space. The Riesz representation theorem says that the mapping

$$\Phi(x)(y) = \langle y, x \rangle, \qquad \Phi : X \to X^*,$$

is an isometric isomorphism. Let U be a nonempty open subset of X and let $f : U \to \mathbb{R}$ be differentiable, with derivative $f' : U \to \mathscr{L}(X; \mathbb{R}) = X^*$.[1] The **gradient of** f is the function $\operatorname{grad} f : U \to X$ defined by

$$\operatorname{grad} f = \Phi^{-1} \circ f'.$$

Thus, for $x \in U$, $\operatorname{grad} f(x)$ is the unique element of X satisfying

$$\langle \operatorname{grad} f(x), y \rangle = f'(x)(y), \qquad y \in X. \tag{1}$$

Because $\Phi^{-1} : X^* \to X$ is continuous, if $f \in C^1(U; \mathbb{R})$ then $\operatorname{grad} f \in C(U; X)$, being a composition of two continuous functions.

For example, let T be a bounded self-adjoint operator on X and define $f : X \to \mathbb{R}$ by

$$f(x) = \frac{1}{2} \langle Tx, x \rangle, \qquad x \in X.$$

For $x, h \in X$,

$$f(x+h) - f(x) = \frac{1}{2} \langle Tx, h \rangle + \frac{1}{2} \langle Th, x \rangle + \frac{1}{2} \langle Th, h \rangle = \langle Tx, h \rangle + \frac{1}{2} \langle Th, h \rangle.$$

Thus

$$|f(x+h) - f(x) - \langle Tx, h \rangle| = \frac{1}{2} |\langle Th, h \rangle| \leq \frac{1}{2} \|T\| \|h\|^2 = o(\|h\|),$$

which shows that f is differentiable at h, with $f'(x)(y) = \langle Tx, y \rangle$. Thus by (1), $\operatorname{grad} f = Tx$.

[1] See http://individual.utoronto.ca/jordanbell/notes/weaksymplectic.pdf, §3.

For example, let $T \in \mathscr{L}(X;X)$, let $h \in X$, and define $f : X \to \mathbb{R}$ by

$$f(x) = \frac{1}{2} \|Tx - h\|^2, \qquad x \in X.$$

We calculate that

$$\operatorname{grad} f(x) = T^*Tx - T^*h, \qquad x \in X.$$

For $x_0 \in X$, define

$$\phi(t) = \exp(-tT^*T)x_0 + \int_0^t \exp(-(t-s)T^*T)T^*h ds, \qquad t \geq 0.$$

It is proved[2] that ϕ satisfies

$$\phi'(t) = -(\operatorname{grad} f)(\phi(t)), \qquad \phi(0) = x_0.$$

For a function $F : X \to X$, we say that F **is** L **Lipschitz** if

$$\|F(x) - F(y)\| \leq L \|x - y\|, \qquad x, y \in X.$$

The following is a useful inequality for functions whose gradients are Lipschitz.[3]

Lemma 1. *If $f : X \to \mathbb{R}$ is differentiable and $\operatorname{grad} f : X \to X$ is L Lipschitz, then*

$$f(y) \leq f(x) + \langle \operatorname{grad} f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2, \qquad x, y \in X.$$

Proof. Let $h = y - x$ and define $g : [0,1] \to \mathbb{R}$ by $g(t) = f(x + th)$. By the chain rule, for $0 < t < 1$,

$$g'(t) = f'(x + th)(h) = \langle \operatorname{grad} f(x + th), h \rangle.$$

Thus by the fundamental theorem of calculus,

$$\int_0^1 \langle \operatorname{grad} f(x + th), h \rangle \, dt = \int_0^1 g'(t)dt = g(1) - g(0) = f(x+h) - f(x) = f(y) - f(x),$$

and so, using the Cauchy-Schwarz inequality and the fact that $\operatorname{grad} f$ is L Lip-

[2] cf. J.W. Neuberger, *A Sequence of Problems on Semigroups*, p. 51, Problem 195.

[3] Juan Peypouquet, *Convex Optimization in Normed Spaces: Theory, Methods and Examples*, p. 15, Lemma 1.30.

schitz,

$$\begin{aligned}
f(y)-f(x) &= \int_0^1 \langle \operatorname{grad} f(x+th) - \operatorname{grad} f(x) + \operatorname{grad} f(x), h\rangle \, dt \\
&= \langle \operatorname{grad} f(x), h\rangle \, dt + \int_0^1 \langle \operatorname{grad} f(x+th) - \operatorname{grad} f(x), h\rangle \, dt \\
&\le \langle \operatorname{grad} f(x), h\rangle + \int_0^1 \|\operatorname{grad} f(x+th) - \operatorname{grad} f(x)\| \, \|h\| \, dt \\
&\le \langle \operatorname{grad} f(x), h\rangle + \int_0^1 L \, \|th\| \, \|h\| \, dt \\
&= \langle \operatorname{grad} f(x), y-x\rangle + \frac{L}{2} \|y-x\|^2 ,
\end{aligned}$$

proving the claim. □

2 Hessians

Let U be a nonempty open subset of X. We prove that if a function is C^2 then its gradient is C^1.[4]

Theorem 2. *Let U be an open subset of X. If $f \in C^2(U;\mathbb{R})$, then $\operatorname{grad} f \in C^1(U;X)$, and*

$$f''(x)(u)(v) = \langle v, (\operatorname{grad} f)'(x)(u)\rangle , \qquad x \in U, \quad u,v \in X. \tag{2}$$

Proof. That f is C^2 means that $f' : U \to X^*$ is C^1. That is, for all $x \in U$, the map $f' : U \to X^*$ is continuous at x, there is $f''(x) \in \mathscr{L}(X;X^*)$ such that

$$\|f'(x+h) - f'(x) - f''(x)(h)\| = o(\|h\|), \tag{3}$$

as $h \to 0$, and the map $x \mapsto f''(x)$ is continuous $U \to \mathscr{L}(X;X^*)$.

Let $x \in U$ and let $h \in X$. Define $\phi_h \in X^*$ by

$$\phi_h(v) = f''(x)(h)(v), \qquad v \in X.$$

Define $\nu_x(h) = \Phi^{-1}(\phi_h) \in X$, thus

$$f''(x)(h)(v) = \langle v, \nu_x(h)\rangle , \qquad v \in X.$$

It is straightforward that ν_x is linear. Because Φ is an isometric isomorphism,

$$\|\nu_x(h)\| = \|\phi_h\| = \sup_{\|v\|\le 1} |\phi_h(v)| = \sup_{\|v\|\le 1} |f''(x)(h)(v)| \le \|f''(x)\| \, \|h\| ,$$

where $(u,v) \mapsto f''(x)(u)(v)$ is a bilinear form, with

$$\|f''(x)\| = \sup_{\|u\|\le 1, \|v\|\le 1} |f''(x)(u)(v)|,$$

[4] Rodney Coleman, *Calculus on Normed Vector Spaces*, p. 139, Theorem 6.5.

showing that $\nu_x : X \to X$ is a bounded linear operator with $\|\nu_x\| \leq \|f''(x)\|$. For h such that $x+h \in U$ and for $v \in X$,

$$(f'(x+h) - f'(x) - f''(x)(h))(v) = \langle v, \operatorname{grad} f(x+h) - \operatorname{grad} f(x) - \nu_x(h) \rangle \,,$$

so

$$\begin{aligned}\|f'(x+h) - f'(x) - f''(x)(h)\| &= \sup_{\|v\| \leq 1} |\, \langle v, \operatorname{grad} f(x+h) - \operatorname{grad} f(x) - \nu_x(h) \rangle \,| \\ &= \|\operatorname{grad} f(x+h) - \operatorname{grad} f(x) - \nu_x(h)\| \,.\end{aligned}$$

Thus by (3),

$$\|\operatorname{grad} f(x+h) - \operatorname{grad} f(x) - \nu_x(h)\| = o(\|h\|)$$

as $h \to 0$, and because $\nu_x \in \mathscr{L}(X;X)$, this means that $\operatorname{grad} f : U \to X$ is differentiable at x, with $(\operatorname{grad} f)'(x) = \nu_x$. It remains to prove that $x \mapsto \nu_x$ is continuous $U \to \mathscr{L}(X;X)$, namely that $(\operatorname{grad} f)'$ is continuous. For $x \in U$ and for h with $x+h \in U$,

$$\begin{aligned}\|\nu_{x+h} - \nu_x\| &= \sup_{\|u\| \leq 1} \|\nu_{x+h}(u) - \nu_x(u)\| \\ &= \sup_{\|u\| \leq 1} \sup_{\|v\| \leq 1} |\, \langle v, \nu_{x+h}(u) - \nu_x(u) \rangle \,| \\ &= \sup_{\|u\| \leq 1} \sup_{\|v\| \leq 1} |f''(x+h)(u)(v) - f''(x)(u)(v)| \\ &= \|f''(x+h) - f''(x)\| \,,\end{aligned}$$

and because f'' is continuous on U we get that $x \mapsto \nu_x$ is continuous on U, completing the proof. $\square$

If $f \in C^2(U;\mathbb{R})$, we proved in the above theorem that $\operatorname{grad} f \in C^1(U;X)$. We call the derivative of $\operatorname{grad} f$ the **Hessian of** f,[5]

$$\operatorname{Hess} f = (\operatorname{grad} f)', \qquad U \to \mathscr{L}(X;X),$$

and (2) then reads

$$f''(x)(u)(v) = \langle v, \operatorname{Hess} f(x)(u) \rangle \,, \qquad x \in U, \quad u, v \in X.$$

Furthermore, it is a fact that if $f \in C^2(U;\mathbb{R})$, then for each $x \in U$, the bilinear form

$$(u,v) \mapsto f''(x)(u)(v)$$

is symmetric.[6] Thus, for $x \in U$ and $u, v \in X$,

$$\langle v, \operatorname{Hess} f(x)(u) \rangle = \langle u, \operatorname{Hess} f(x)(v) \rangle \,.$$

[5] cf. R. A. Tapia, *The differentiation and integration of nonlinear operators*, pp. 45–101, in *Nonlinear Functional Analysis and Applications* (Louis B. Rall, ed.)

[6] Serge Lang, *Real and Functional Analysis*, third ed., p. 344, Theorem 5.3.

Now, using that $\langle \cdot, \cdot \rangle$ is symmetric as X is a real Hilbert space, $(\operatorname{Hess} f(x))^* \in \mathscr{L}(X; X)$ satisfies

$$\langle u, \operatorname{Hess} f(x)(v) \rangle = \langle (\operatorname{Hess} f(x))^* u, v \rangle = \langle v, (\operatorname{Hess} f(x))^* u \rangle .$$

so

$$\langle v, \operatorname{Hess} f(x)(u) \rangle = \langle v, (\operatorname{Hess} f(x))^* u \rangle .$$

Because this is true for all v we have $\operatorname{Hess} f(x)(u) = (\operatorname{Hess} f(x))^* u$, and because this is true for all u we have $\operatorname{Hess} f(x) = (\operatorname{Hess} f(x))^*$, i.e. $\operatorname{Hess} f(x)$ is self-adjoint.

Theorem 3. *If U is an open subset of X and $f \in C^2(U; \mathbb{R})$, then for each $x \in U$ it is the case that $\operatorname{Hess} f(x) \in \mathscr{L}(X; X)$ is self-adjoint.*

3 Critical points

For an open set U in X for $k \geq 1$, and for $f \in C^{k+2}(U; \mathbb{R})$, we say that $x_0 \in U$ is a **critical point of** f if $f'(x_0) = 0$. If x_0 is a critical point of f, let we say that x_0 is a **nondegenerate critical point of** f if $\operatorname{Hess} f(x_0) \in \mathscr{L}(X; X)$ is invertible. The **Morse-Palais lemma**[7] states that if $f \in C^{k+2}(U; \mathbb{R})$ with $k \geq 1$, $f(0) = 0$, and 0 is a nondegenerate critical point of f, then there is some open subset V of U with $0 \in V$ and a C^k diffeomorphism $\phi : V \to V$, $\phi(0) = 0$, such that

$$f(x) = \frac{1}{2} \langle \operatorname{Hess} f(0)(\phi(x)), \phi(x) \rangle , \qquad x \in V.$$

If x is a critical point of a differentiable function $f : U \to \mathbb{R}$, we call $f(x)$ a **critical value of** f. If $k \geq n$ and $f \in C^k(\mathbb{R}^n; \mathbb{R})$, **Sard's theorem** tells us that the set of critical values of f has Lebesgue measure 0 and is meager.

For Banach spaces Y and Z, a **Fredholm operator**[8] is a bounded linear operator $T : Y \to Z$ such that (i) $\alpha(T) = \dim \ker T < \infty$, (ii) $T(Y)$ is a closed subset of Z, and (iii) $\beta(T) = \dim \ker T^* < \infty$. The **index** of a Fredholm operator T is

$$\operatorname{ind} T = \alpha(T) - \beta(T).$$

For a differentiable function $f : U \to \mathbb{R}$, U an open subset of X, and for $x \in U$, $f'(x) \in \mathscr{L}(X; \mathbb{R}) = X^*$. $f'(x)$ is a Fredholm operator if and only if $\dim \ker f'(x) < \infty$. For U a connected open subset of X and for $f \in C^1(U; \mathbb{R})$, we call f a **Fredholm map** if $f'(x)$ is a Fredholm operator for each $x \in U$. It

[7] Serge Lang, *Differential and Riemannian Manifolds*, p. 182, chapter VII, Theorem 5.1; Kung-ching Chang, *Infinite Dimensional Morse Theory and Multiple Solution Problems*, p. 33, Theorem 4.1; André Avez, *Calcul différentiel*, p. 87, §3; N. A. Bobylev, S. V. Emel'yanov, and S. K. Korovin, *Geometrical Methods in Variational Problems*, p. 360, Theorem 5.5.2; Hajime Urakawa, *Calculus of Variations and Harmonic Maps*, p. 87, chapter 3, §1, Theorem 1.10; Jean-Pierre Aubin and Ivar Ekeland, *Applied Nonlinear Analysis*, p. 52, Theorem 8; Melvyn S. Berger, *Nonlinearity and Functional Analysis: Lectures on Nonlinear Problems in Mathemtical Analysis*, p. 355, Theorem 6.5.4.

[8] Martin Schechter, *Principles of Functional Analysis*, second ed., chapter 5.

is a fact that $\operatorname{ind} f'(x) = \operatorname{ind} f'(y)$ for all $x, y \in U$, using that U is connected. We denote this common value by $\operatorname{ind} f$. A generalization of Sard's theorem by Smale here tells us that if X is separable, U is a connected open subset of X, $f \in C^k(U; \mathbb{R})$ is a Fredholm map, and

$$k > \max\{\operatorname{ind} f, 0\},$$

then the set of critical values of f is meager.[9]

A function $f \in C^1(X; \mathbb{R})$ is said to **satisfy the Palais-Smale condition** if (u_k) is a sequence in X such that (i) $\{f(u_k)\}$ is a bounded subset of $\mathbb{R}$ and (ii) $\operatorname{grad} f(u_k) \to 0$, then $\{u_k\}$ is a precompact subset of X: every subsequence of (u_k) itself has a Cauchy subsequence.

Often when speaking about ordinary differential equations in $\mathbb{R}^d$, we deal with differentiable functions whose derivatives are locally Lipschitz. $\mathbb{R}^d$ has the Heine-Borel property: a subset K of $\mathbb{R}^d$ is compact if and only if K is closed and bounded. In fact no infinite dimensional Banach space has the Heine-Borel property.[10] Thus a locally Lipschitz function need not be Lipschitz on a bounded subset of X. (On a compact set, the set is covered by balls on which the function is Lipschitz, and then the function is Lipschitz on the compact set with Lipschitz constant equal to the maximum of finitely many Lipschitz constants on the balls.) We denote by $\mathcal{C}$ the set of function $f : X \to \mathbb{R}$ that are differentiable and such that for each bounded subset A of X, the restriction of $\operatorname{grad} f$ to A is Lipschitz.

The **mountain pass theorem**[11] states that if (i) $I \in \mathcal{C}$, (ii) I satisfies the Palais-Smale condition, (iii) $I(0) = 0$, (iv) there are $r, a > 0$ such that $I(u) \geq a$ when $\|u\| = r$, and (v) there is some $v \in X$ satisfying $\|v\| > r$ and $I(v) \leq 0$, then

$$\inf_{g \in \Gamma_v} \sup_{0 \leq t \leq 1} (I \circ g)(t)$$

is a critical value of I, where

$$\Gamma_v = \{g \in C([0, 1]; X) : g(0) = 0, g(1) = v\}.$$

4 Convexity

We prove that a critical point of a differentiable convex function on an open convex set is a minimum.[12]

[9] Eberhard Zeidler, *Nonlinear Functional Analysis and its Applications, IV: Applications to Mathematical Physics*, p. 829, Theorem 78.A; Melvyn S. Berger, *Nonlinearity and Functional Analysis: Lectures on Nonlinear Problems in Mathematical Analysis*, p. 125, Theorem 3.1.45.

[10] Some Fréchet spaces have the Heine-Borel property, like the space of holomorphic functions on the open unit disc, which is what Montel's theorem says: `http://individual.utoronto.ca/jordanbell/notes/holomorphic.pdf`, Theorem 13.

[11] Lawrence C. Evans, *Partial Differential Equations*, p. 480, Theorem 2; Antonio Ambrosetti and David Arcoya lvarez, *An Introduction to Nonlinear Functional Analysis and Elliptic Problems*, p. 48, §5.3.

[12] N. A. Bobylev, S. V. Emel'yanov, and S. K. Korovin, *Geometrical Methods in Variational Problems*, p. 39, Theorem 2.1.4.

Theorem 4. *If A is an open convex set, $f : A \to \mathbb{R}$ is differentiable and convex, and $x_0 \in A$ is a critical point of f, then $f(x_0) \leq f(x)$ for all $x \in A$.*

Proof. Because f is convex, for $0 < t < 1$,

$$f(tx + (1-t)x_0) \leq tf(x) + (1-t)f(x_0),$$

i.e.

$$\frac{f(x_0 + t(x - x_0)) - f(x_0)}{t} \leq f(x) - f(x_0).$$

Taking $t \to 0$,

$$f'(x_0)(x - x_0) \leq f(x) - f(x_0),$$

and because x_0 is a critical point,

$$0 \leq f(x) - f(x_0),$$

i.e. $f(x_0) \leq f(x)$. □

We establish equivalent conditions for a differentiable function to be convex.[13]

Theorem 5. *If A is an open convex subset of X and $f : A \to \mathbb{R}$ is differentiable, then the following are equivalent:*

1. *f is convex.*
2. *$f(y) \geq f(x) + \langle \operatorname{grad} f(x), y - x \rangle$, $x, y \in A$.*
3. *$\langle \operatorname{grad} f(x) - \operatorname{grad} f(y), x - y \rangle \geq 0$, $x, y \in A$.*

Proof. Suppose (1). For $x, y \in A$ and $0 < t < 1$, that f is convex means $f(ty + (1-t)x) \leq tf(y) + (1-t)f(x)$, i.e.

$$\frac{f(x + t(y - x)) - f(x)}{t} \leq f(y) - f(x),$$

and taking $t \to 0$ yields

$$f'(x)(y - x) \leq f(y) - f(x),$$

i.e.

$$\langle \operatorname{grad} f(x), y - x \rangle \leq f(y) - f(x).$$

Suppose (2) and let $x, y \in A$, for which

$$\langle \operatorname{grad} f(x), y - x \rangle \leq f(y) - f(x), \qquad \langle \operatorname{grad} f(y), x - y \rangle \leq f(x) - f(y).$$

[13] Juan Peypouquet, *Convex Optimization in Normed Spaces: Theory, Methods and Examples*, p. 38, Proposition 3.10.

Adding these inequalities,

$$\langle \operatorname{grad} f(x), y-x \rangle - \langle \operatorname{grad} f(y), y-x \rangle \le 0.$$

Suppose (3), let $x, y \in A$, and define $\phi : [0,1] \to \mathbb{R}$ by

$$\phi(t) = f(tx + (1-t)y) - tf(x) - (1-t)f(y).$$

$\phi(0) = 0$ and $\phi(1) = 0$, and for $0 < t < 1$, using the chain rule gives

$$\begin{aligned} \phi'(t) &= f'(tx + (1-t)y)(x-y) - f(x) + f(y) \\ &= \langle \operatorname{grad} f(tx + (1-t)y), x-y \rangle - f(x) + f(y). \end{aligned}$$

Let $0 < s < t < 1$, let $u = sx + (1-s)y$ and $v = tx + (1-t)y$, which both belong to A because A is convex, and so the above reads

$$\phi'(s) = \langle \operatorname{grad} f(u), x-y \rangle - f(x) + f(y), \qquad \phi'(t) = \langle \operatorname{grad} f(v), x-y \rangle - f(x) + f(y),$$

so

$$\phi'(s) - \phi'(t) = \langle \operatorname{grad} f(u) - \operatorname{grad} f(v), x-y \rangle.$$

And

$$(s-t)(x-y) = u - y - (v-y) = u - v,$$

so

$$\phi'(s) - \phi'(t) = \frac{1}{s-t} \langle \operatorname{grad} f(u) - \operatorname{grad} f(v), u-v \rangle.$$

But (3) tells us

$$\langle \operatorname{grad} f(u) - \operatorname{grad} f(v), u-v \rangle \ge 0,$$

so, as $s - t < 0$,

$$\phi'(s) - \phi'(t) \le 0,$$

showing that ϕ' is nondecreasing. On the other hand, because $\phi(0) = 0$ and $\phi(1) = 0$, by the mean value theorem there is some $0 < t_0 < 1$ for which $\phi'(t_0) = 0$. Therefore, because ϕ' is nondecreasing it holds that

$$\phi'(t) \le 0, \qquad 0 \le t \le t_0,$$

and

$$\phi'(t) \ge 0, \qquad t_0 \le t \le 1.$$

That is, ϕ is nonincreasing on $[0, t_0]$, and with $\phi(0) = 0$ this yields $\phi(t) \le 0$ for $t \in [0, t_0]$, and ϕ is nondecreasing on $[t_0, 1]$, and with $\phi(1) = 0$ this yields $\phi(t) \le 0$ for $t \in [t_0, 1]$. Therefore $\phi(t) \le 0$ for $t \in [0,1]$, which means that

$$f(tx + (1-t)y) - tf(x) - (1-t)f(y) \le 0, \qquad 0 \le t \le 1,$$

showing that f is convex. $\square$

Theorem 6. *If A is an open convex subset of X and $f : A \to \mathbb{R}$ is twice differentiable, then the following are equivalent:*

1. *f is convex.*
2. $\langle \operatorname{Hess} f(x)(v), v \rangle \geq 0$, *$x \in A$, $v \in X$.*

Proof. Suppose (1) and let $x \in A$. From Theorem 5, $v \in X$ and for $t > 0$ with which $x + tv \in A$,

$$\langle \operatorname{grad} f(x + tv) - \operatorname{grad} f(x), tv \rangle \geq 0,$$

i.e.

$$\frac{f'(x + tv)(v) - f'(x)(v)}{t} \geq 0.$$

Taking $t \to 0$,

$$f''(x)(v)(v) \geq 0,$$

i.e.

$$\langle \operatorname{Hess} f(x)(v), v \rangle \geq 0.$$

Suppose (2), let $x, y \in A$ and define $\phi : [0, 1] \to \mathbb{R}$ by

$$\phi(t) = f(tx + (1 - t)y) - tf(x) - (1 - t)f(y).$$

Applying the chain rule, for $0 < t < 1$,

$$\phi''(t) = f''(tx + (1 - t)y)(x - y)(x - y),$$

i.e.

$$\phi''(t) = \langle \operatorname{Hess} f(tx + (1 - t)y)(x - y), x - y \rangle \geq 0,$$

showing that ϕ' is nondecreasing. In the proof of Theorem 5 we deduced from ϕ' being nondecreasing and satisfying $\phi(0) = 0$, $\phi(1) = 0$, that f is convex, and the same reasoning yields here that f is convex. $\square$

We call a function $F : X \to X$ β **co-coercive** if

$$\langle F(x) - F(y), x - y \rangle \geq \beta \left\| F(x) - F(y) \right\|^2.$$

We prove conditions under which the gradient of a differentiable convex function is co-coercive.[14]

Theorem 7 (Baillon-Haddad theorem). *Let $f : X \to \mathbb{R}$ be differentiable and convex and let $L > 0$. Then $\operatorname{grad} f$ is L Lipschitz if and only if $\operatorname{grad} f$ is $\frac{1}{L}$ co-coercive.*

Proof. Suppose that $\operatorname{grad} f$ is L Lipschitz and for $x \in X$, define $h_x : X \to \mathbb{R}$ by

$$h_x(y) = f(y) - f'(x)(y) = f(y) - \langle \operatorname{grad} f(x), y \rangle.$$

[14] Juan Peypouquet, *Convex Optimization in Normed Spaces: Theory, Methods and Examples*, p. 40, Theorem 3.13.

For $y, z \in X$ and $0 < t < 1$, because f is convex,

$$\begin{aligned} h_x(tz + (1-t)y) &= f(tz + (1-t)y) - \langle \operatorname{grad} f(x), tz + (1-t)y \rangle \\ &\leq tf(z) + (1-t)f(y) - \langle \operatorname{grad} f(x), tz + (1-t)y \rangle \\ &= th_x(z) + (1-t)h_x(y), \end{aligned}$$

showing that h_x is convex. For $y, z \in X$,[15]

$$h_x'(y)(z) = f'(y)(z) - f'(x)(z),$$

and in particular $\operatorname{grad} h_x(x) = 0$. Thus by Theorem 4,

$$h_x(x) \leq h_x(y), \qquad y \in X. \tag{4}$$

For $x, y, z \in X$, by Lemma 1,

$$f(z) \leq f(x) + \langle \operatorname{grad} f(x), z - x \rangle + \frac{L}{2} \|z - x\|^2,$$

so

$$h_y(z) \leq f(x) - \langle \operatorname{grad} f(y), z \rangle + \langle \operatorname{grad} f(x), z - x \rangle + \frac{L}{2} \|z - x\|^2,$$

i.e.

$$h_y(z) \leq h_x(x) + \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), z \rangle + \frac{L}{2} \|z - x\|^2,$$

and applying (4),

$$h_y(y) \leq h_x(x) + \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), z \rangle + \frac{L}{2} \|z - x\|^2. \tag{5}$$

Now,

$$\|\operatorname{grad} f(x) - \operatorname{grad} f(y)\| = \sup_{\|v\| \leq 1} \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), v \rangle$$

so for each $\epsilon > 0$ there is some $v_\epsilon \in X$ with $\|v_\epsilon\| \leq 1$ and

$$\langle \operatorname{grad} f(x) - \operatorname{grad} f(y), v_\epsilon \rangle \geq \|\operatorname{grad} f(x) - \operatorname{grad} f(y)\| - \epsilon.$$

Let $R = \frac{\|\operatorname{grad} f(x) - \operatorname{grad} f(y)\|}{L}$, and applying (5) with $z = x - Rv_\epsilon$ yields

$$\begin{aligned} h_y(y) &\leq h_x(x) + \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), x - Rv_\epsilon \rangle + \frac{L}{2} \|Rv_\epsilon\|^2 \\ &= h_x(x) + \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), x \rangle - R \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), v_\epsilon \rangle \\ &+ \frac{1}{2L} \|\operatorname{grad} f(x) - \operatorname{grad} f(y)\|^2 \|v_\epsilon\|^2 \\ &\leq h_x(x) + \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), x \rangle - R \|\operatorname{grad} f(x) - \operatorname{grad} f(y)\| + R\epsilon \\ &+ \frac{1}{2L} \|\operatorname{grad} f(x) - \operatorname{grad} f(y)\|^2 \\ &= h_x(x) + \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), x \rangle - \frac{1}{2L} \|\operatorname{grad} f(x) - \operatorname{grad} f(y)\|^2 + R\epsilon. \end{aligned}$$

[15] Henri Cartan, *Differential Calculus*, p. 29, Proposition 2.4.2.

Likewise, because R does not change when x and y are switched,

$$h_x(x) \le h_y(y) + \langle \operatorname{grad} f(y) - \operatorname{grad} f(x), y \rangle - \frac{1}{2L} \left\| \operatorname{grad} f(y) - \operatorname{grad} f(x) \right\|^2 + R\epsilon.$$

Adding these inequalities,

$$\begin{aligned} 0 \le &\langle \operatorname{grad} f(x) - \operatorname{grad} f(y), x \rangle + \langle \operatorname{grad} f(y) - \operatorname{grad} f(x), y \rangle \\ &- \frac{1}{L} \left\| \operatorname{grad} f(x) - \operatorname{grad} f(y) \right\|^2 + 2R\epsilon, \end{aligned}$$

i.e.

$$\frac{1}{L} \left\| \operatorname{grad} f(x) - \operatorname{grad} f(y) \right\|^2 \le \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), x - y \rangle + 2R\epsilon.$$

This is true for all $\epsilon > 0$, so

$$\frac{1}{L} \left\| \operatorname{grad} f(x) - \operatorname{grad} f(y) \right\|^2 \le \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), x - y \rangle,$$

showing that $\operatorname{grad} f$ is $\frac{1}{L}$ co-coercive.

Suppose that $\operatorname{grad} f$ is $\frac{1}{L}$ co-coercive and let $x, y \in X$. Then applying the Cauchy-Schwarz inequality,

$$\begin{aligned} \left\| \operatorname{grad} f(x) - \operatorname{grad} f(y) \right\|^2 &\le L \langle \operatorname{grad} f(x) - \operatorname{grad} f(y), x - y \rangle \\ &\le L \left\| \operatorname{grad} f(x) - \operatorname{grad} f(y) \right\| \left\| x - y \right\|. \end{aligned}$$

If $\left\| \operatorname{grad} f(x) - \operatorname{grad} f(y) \right\| = 0$ then certainly $\left\| \operatorname{grad} f(x) - \operatorname{grad} f(y) \right\| \le L \left\| x - y \right\|$. Otherwise, dividing by $\left\| \operatorname{grad} f(x) - \operatorname{grad} f(y) \right\|$ gives

$$\left\| \operatorname{grad} f(x) - \operatorname{grad} f(y) \right\| \le L \left\| x - y \right\|,$$

showing that $\operatorname{grad} f$ is L Lipschitz. $\square$