6 Metric Gradient

Goal

Metric gradient, sign gradient descent, coordinate gradient descent, convergence, metric projection

Alert 6.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.
This note is likely to be updated again soon.

Definition 6.2: Problem

The problem we study in this lecture is the following:

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) = f(w_1, \dots, w_d),$$

where f is a $\mathsf{L} = \mathsf{L}^{[1]}$-smooth function w.r.t. a general norm $\|\cdot\|$. For simplicity, we do not consider any constraints on $\mathbf{w}$. Below, we will learn how to adapt the gradient descent Algorithm 1.4 to any norm.

Example 6.3: Distributed learning through gradient compression

Let us consider the typical formulation in ML:

$$f(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{w}; \mathcal{D}_i),$$

where $\mathcal{D}_i$ is some data for the i-th component function f_i. For instance, f_i could represent the contribution from a server or user in a different location or from a study in a different time or even from a different processor in the same system. The gradient of f is obviously

$$\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{w}; \mathcal{D}_i).$$

Since the component gradients $\nabla f_i(\mathbf{w}; \mathcal{D}_i)$ are computed separately, they need to be communicated to a central server who does the aggregation. The communication cost is proportional to d and the numerical precision (i.e. the number of bits) we use to represent each entry in the gradient. When d is large, it makes sense to *compress* the component gradients before communicating to the server. Sometimes limitation of hardware also makes such compression inevitable.

A common practice in resource limited distributed learning is to take the sign (phase) of the component gradients while completely discard the magnitude information, see for instance Bernstein et al. (2018, 2019) and Balles and Hennig (2018) for some recent applications. The obvious question to ask is:

What is the impact of compression on the convergence (rate) of a gradient algorithm?

Bernstein, J., Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar (2018). "signSGD: Compressed Optimisation for Non-Convex Problems". In: *Proceedings of the 35th International Conference on Machine Learning*, pp. 560–569.
Bernstein, J., J. Zhao, K. Azizzadenesheli, and A. Anandkumar (2019). "signSGD with Majority Vote is Communication Efficient and Fault Tolerant". In: *International Conference on Learning Representations*.
Balles, L. and P. Hennig (2018). "Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients". In: *ICML*.

Definition 6.4: Metric gradient (Golomb and Tapia, 1972)

Let $f : \mathsf{V} \to \mathbb{R}$ be differentiable. Following Golomb and Tapia (1972) we define the metric gradient as:

$$\blacktriangledown f(\mathbf{w}) := \operatorname*{argmax}_{\|\mathbf{d}\| \le \|\nabla f(\mathbf{w})\|_\circ} \langle \mathbf{d}; \nabla f(\mathbf{w}) \rangle, \quad \overline{\nabla} f : \mathsf{V} \to \mathsf{V}, \ \mathbf{w} \mapsto \overline{\nabla} f(\mathbf{w}) \in \blacktriangledown f(\mathbf{w}). \tag{6.1}$$

Note that unlike the gradient ∇f which is a topological quantity, the metric gradient $\overline{\nabla} f$ *does* depend on the norm. It follows from definition that

$$\langle \overline{\nabla} f(\mathbf{w}); \nabla f(\mathbf{w}) \rangle = \| \overline{\nabla} f(\mathbf{w}) \|^2 = \|\nabla f(\mathbf{w})\|_\circ^2.$$

Golomb, M. and R. A. Tapia (1972). "The metric gradient in normed linear spaces". *Numerische Mathematik*, vol. 20, pp. 115–124.

Definition 6.5: Duality mapping

Let $\mathsf{q} = \frac{1}{2}\|\cdot\|^2$ be the "quadratic" function w.r.t. a general norm. It is obviously convex, and we call its subdifferential the duality mapping

$$\mathsf{J} = \partial \mathsf{q} : \mathsf{V} \to 2^{\mathsf{V}^*}, \quad \mathsf{j} : \mathsf{V} \to \mathsf{V}^*, \ \mathbf{w} \mapsto \mathsf{j}(\mathbf{w}) \in \mathsf{J}(\mathbf{w}),$$

where j is an arbitrary single-valued selection of J. We have the following properties of the duality mapping:

$$\langle \mathbf{w}; \mathsf{j}(\mathbf{w}) \rangle = \|\mathbf{w}\|^2 = \|\mathsf{j}(\mathbf{w})\|_\circ^2.$$

We can now also define the metric gradient as

$$\blacktriangledown f = \mathsf{J}^{-1}(\nabla f), \quad \overline{\nabla} f = \mathsf{j}^{-1}(\nabla f),$$

provided that the inverse duality map J^{-1} exists (e.g. when the space V is reflexive).

Definition 6.6: Steepest descent (Kantorovich, 1945)

Another way to recognize the metric gradient is through Kantorovich's steepest descent. Fixing the current iterate $\mathbf{w}_t$, we look for a direction $\mathbf{d}$ such that the univariate function

$$\eta \mapsto h(\eta) := f(\mathbf{w}_t - \eta \mathbf{d})$$

decreases steepest. Kantorovich (1945) proposed to find the direction $\mathbf{d}$ through the following subproblem:

$$\operatorname*{argmin}_{\mathbf{d} \neq \mathbf{0}} \frac{h'(\eta)|_{\eta=0}}{\|\mathbf{d}\|} = \frac{-\langle \mathbf{d}; \nabla f(\mathbf{w}_t) \rangle}{\|\mathbf{d}\|} \implies \mathbf{d} = \frac{\overline{\nabla} f(\mathbf{w}_t)}{\| \overline{\nabla} f(\mathbf{w}_t)\|} = \frac{\overline{\nabla} f(\mathbf{w}_t)}{\|\nabla f(\mathbf{w}_t)\|_\circ},$$

which is exactly the normalized metric gradient! Being an expert in functional analysis and a pioneer in numerical analysis, Kantorovich clearly recognized the generality in the choice of the norm, although many of his examples are still in the Euclidean/Hilbert setting.

Kantorovich, L. V. (1945). "On an effective method of solving extremal problems for quadratic functionals". *Soviet Mathematics Doklady*, vol. 48, no. 7, pp. 595–600.

Example 6.7: ℓ_p norm metric gradient

Let $\mathsf{V} = \mathbb{R}^d$ be equipped with the ℓ_p norm, whose dual is ℓ_q norm with $1/p + 1/q = 1$ (see Definition 0.10). We easily compute

$$\blacktriangledown f(\mathbf{w}) := \left[\operatorname*{argmax}_{\|\mathbf{z}\|_p \le \|\nabla f(\mathbf{w})\|_q} \langle \mathbf{z}; \nabla f(\mathbf{w}) \rangle \right] = \|\nabla f(\mathbf{w})\|_q^{1-q/p} \cdot \operatorname{sign}(\nabla f(\mathbf{w})) \cdot |\nabla f(\mathbf{w})|^{q/p}.$$

In particular,

- when $p = q = 2$, we have $\blacktriangledown f = \nabla f$;
- when $p = 1, q = \infty$, we have $\blacktriangledown f = \operatorname{conv}\{\nabla_j f \cdot \mathbf{e}_j : |\nabla_j f| = \|\nabla f\|_\infty\}$;
- when $p = \infty, q = 1$, we have $\blacktriangledown f = \operatorname{conv}\{\|\nabla f\|_1 \cdot \operatorname{sign}(\nabla f)\}$, where $\operatorname{sign}(0) \in [-1, 1]$.

We verify that the metric gradient indeed depends on the norm.

Algorithm 6.8: Metric gradient descent (Golomb and Tapia, 1972)

Algorithm: Metric gradient descent for unconstrained smooth minimization

Input: $\mathbf{w}_0$, norm $\|\cdot\|$

for $t = 0, 1, \ldots$ **do**
  $\mathbf{g}_t \leftarrow \overline{\nabla} f(\mathbf{w}_t)$ `// compute any metric gradient`
  if $\|\mathbf{g}_t\| = 0$ **then**
    break
  choose step size $\eta_t > 0$
  $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta_t \mathbf{g}_t$ `// update`

Golomb, M. and R. A. Tapia (1972). "The metric gradient in normed linear spaces". *Numerische Mathematik*, vol. 20, pp. 115–124.

Example 6.9: Sign gradient descent

If we choose the ℓ_∞ norm in Algorithm 6.8, we obtain the so-called sign gradient descent algorithm, where in each iteration we only update with the sign of gradient:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \|\nabla f(\mathbf{w}_t)\|_1 \cdot \operatorname{sign}(\nabla f(\mathbf{w}_t)).$$

We choose *not* to absorb the scalar $\|\nabla f(\mathbf{w}_t)\|_1$ into the step size, as it reveals the true nature of the sign gradient algorithm: merely a special case of the metric gradient Algorithm 6.8. This algorithm is particularly appealing in distributed and low-resource devices.

Example 6.10: Coordinate gradient descent

If we choose the ℓ_1 norm in Algorithm 6.8, we obtain the so-called greedy coordinate gradient descent algorithm (Southwell, 1935), where in each iteration we only take a gradient step along one (block of) coordinate(s):

$$w_{j,t+1} = w_{j,t} - \eta_t \nabla_j f(\mathbf{w}_t), \quad \text{where} \quad j \text{ is chosen as any index such that } |\nabla_j f| = \|\nabla f\|_\infty.$$

This algorithm is deemed inefficient in theory: one computes the entire gradient (in order to decide the index j) and then discards most of the computational effort in the update (since we only update 1 coordinate)!

An obvious alternative is to update the coordinates cyclically:

$$\begin{aligned}&\textbf{for}\ \ j = 1, \dots, d\\ &\qquad w_j \leftarrow w_j - \eta \nabla_j f(\mathbf{w}).\end{aligned}$$

However, this alternative is also problematic: for most functions we use in practice, computing the gradient vector ∇f costs as much as computing a single component gradient $\nabla_j f$! Thus, the above cyclic coordinate gradient algorithm is d times more expensive than the usual gradient algorithm! However, the catch is that the $\mathsf{L}^{[1]}$-smoothness parameters in coordinate gradient and gradient descent (w.r.t. ℓ_2 norm) are different, and in certain cases the former may still be advantageous.

Indeed, the picture turns around when we randomize our choice of the coordinates (Nesterov, 2012). Interestingly, Nutini et al. (2015) argued that the greedy alternative can be more efficient than the randomized alternative in some settings.

In fact, if we are just updating a single coordinate, we might as well go to the extreme:

$$w_j \leftarrow \operatorname*{argmin}_w\ f(w_1, \dots, w_{j-1}, w, w_{j+1}, \dots, w_d).$$

This algorithm, known as alternating minimization, is extremely popular in practice and will be analyzed in a later lecture.

Southwell, R. V. (1935). "Stress-Calculation in Frameworks by the Method of "Systematic Relaxation of Constraints". I and II". *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 151, no. 872, pp. 56–95.

Nesterov, Y. (2012). "Efficiency of Coordinate Descent Methods on Huge-Scale Optimization Problems". *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362.

Nutini, J., M. Schmidt, I. Laradji, M. Friedlander, and H. Koepke (2015). "Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection". In: *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1632–1641.

Remark 6.11: Metric gradient descent as quadratic upper bound minimization

Suppose f is $\mathsf{L} = \mathsf{L}^{[1]}$-smooth w.r.t. a general norm $\|\cdot\|$, then for any $\eta_t \le 1/\mathsf{L}$ and all $\mathbf{w}$:

$$f(\mathbf{w}) \le f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t; \nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|^2.$$

To minimize the "quadratic" upper bound on the right-hand side, we apply polar decomposition:

$$\min_{\lambda \ge 0} \min_{\|\mathbf{w}-\mathbf{w}_t\|=\lambda} f(\mathbf{w}_t) + \langle \mathbf{w} - \mathbf{w}_t; \nabla f(\mathbf{w}_t)\rangle + \tfrac{1}{2\eta_t}\lambda^2 \quad \equiv \quad \min_{\lambda \ge 0} -\lambda \|\nabla f(\mathbf{w}_t)\|_\circ + \tfrac{1}{2\eta_t}\lambda^2.$$

Thus, $\lambda = \eta_t \|\nabla f(\mathbf{w}_t)\|_\circ$ and

$$\mathbf{w} - \mathbf{w}_t = \lambda \frac{-\overline{\nabla} f(\mathbf{w}_t)}{\|\nabla f(\mathbf{w}_t)\|_\circ}, \qquad i.e. \qquad \mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\, \overline{\nabla} f(\mathbf{w}_t).$$

Theorem 6.12: Convergence of metric gradient descent for $\mathsf{L}^{[1]}$-smooth functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be $\mathsf{L} = \mathsf{L}^{[1]}$-smooth w.r.t. a general norm $\|\cdot\|$ and bounded from below (i.e. $f_\star > -\infty$). If the step size $\eta_t \in [\alpha, \frac{2}{\mathsf{L}} - \beta]$ for some $\alpha, \beta > 0$, then the sequence $\{\mathbf{w}_t\}$ generated by Algorithm 6.8 satisfies $\overline{\nabla} f(\mathbf{w}_t) \to \mathbf{0}$. Moreover,

$$\min_{0 \le t \le T-1} \|\overline{\nabla} f(\mathbf{w}_t)\| \le \sqrt{\frac{f(\mathbf{w}_0) - f_\star}{\alpha\beta\mathsf{L}T/2}}.$$

Proof: The proof is literally the same as that of Theorem 1.17. ■

Choosing $\alpha = \beta = \frac{1}{\mathsf{L}}$, the bound reduces to

$$\min_{0\le t\le T-1} \| \nabla f(\mathbf{w}_t)\| \le \sqrt{\frac{2\mathsf{L}[f(\mathbf{w}_0) - f_\star]}{T}}. \tag{6.2}$$

Obviously, the left-hand side depends on the norm and so does the right-hand side (through $\mathsf{L} = \mathsf{L}_{\|\cdot\|}$).

Alert 6.13: Which norm to use?

Suppose we have two norms $\|\mathbf{w}\| \ge \|\!|\mathbf{w}|\!\|$ for all $\mathbf{w}$. Then, clearly

$$\mathsf{L}_{\|\cdot\|} \le \mathsf{L}_{\|\!|\cdot|\!\|}.$$

However, a direct comparison in (6.2) is not obvious: the metric gradient itself changes with the norm and how we measure the metric gradient also changes with the norm. Not to mention the possibility that computing different metric gradients can incur vastly different costs.

Nevertheless, let us do a quick sanity check. Let $\|\cdot\| = \lambda\|\!|\cdot|\!\|$. Then,

$$\mathsf{L}_{\|\cdot\|} \le \tfrac{1}{\lambda^2}\mathsf{L}_{\|\!|\cdot|\!\|}, \qquad \nabla_{\|\cdot\|} f = \tfrac{1}{\lambda^2}\nabla_{\|\!|\cdot|\!\|} f.$$

Plugging into the bound (6.2) we see that λ is canceled out. So, simply scaling a norm cannot and should not change our bound ☺.

Theorem 6.14: Limit points (if any) of metric gradient iterates are critical

Let $f : \mathsf{V} \to \mathbb{R}$ be continuously differentiable, and we make the following choices in Algorithm 6.8: $\exists \alpha \in (0, 1], \beta_t \to 0, \epsilon_t \to 0, \delta_t \to 0$ such that

- *$\|\mathbf{g}_t\| \le \|\nabla f(\mathbf{w}_t)\|_\circ$ and $\langle \mathbf{g}_t; \nabla f(\mathbf{w}_t)\rangle \ge \alpha \cdot \|\nabla f(\mathbf{w}_t)\|_\circ^2 - \delta_t$, i.e. $\mathbf{g}_t$ is an (α, δ_t)-approximate maximizer of (6.1). This relaxation is conceptually important as it ensures the existence of $\mathbf{g}_t$.*
- *$\eta_t = 0$ iff $\nabla f(\mathbf{w}_t) = \mathbf{0}$.*
- *$\forall \eta \in [0, \eta_t]$, $f(\mathbf{w}_t - \eta \mathbf{g}_t) \ge f(\mathbf{w}_t - \eta_t \mathbf{g}_t) - \beta_t \eta$. In particular, setting $\eta = 0$ we see the algorithm is descending: $f(\mathbf{w}_t) \ge f(\mathbf{w}_{t+1})$.*
- *$\langle \mathbf{g}_t; \nabla f(\mathbf{w}_t - \eta_t \mathbf{g}_t)\rangle \le \epsilon_t$.*

Then, any limit point $\mathbf{w}_*$ of the sequence $\{\mathbf{w}_t\}$ is stationary, i.e. $\nabla f(\mathbf{w}_*) = \mathbf{0}$.

Proof: Due to the second condition above, we may assume $\eta_t > 0$ for all t. Suppose to the contrary $\nabla f(\mathbf{w}_*) \ne \mathbf{0}$. Choose ϵ such that $0 < \epsilon < \alpha\|\nabla f(\mathbf{w}_*)\|_\circ^2$. For the subsequence $\mathbf{w}_k := \mathbf{w}_{t_k} \to \mathbf{w}_*$ we may assume w.l.o.g. that $\alpha\|\nabla f(\mathbf{w}_k)\|_\circ^2 - \delta_k > \epsilon > \epsilon_k$. Let

$$s_k = \inf\{\eta \in [0, \eta_k] : \langle \mathbf{g}_k; \nabla f(\mathbf{w}_k - \eta \mathbf{g}_k)\rangle \le \epsilon\}. \tag{6.3}$$

Since $\langle \mathbf{g}_k; \nabla f(\mathbf{w}_k)\rangle \ge \alpha\|\nabla f(\mathbf{w}_k)\|_\circ^2 - \delta_k > \epsilon$, using continuity we know $s_k > 0$. The last condition implies $\langle \mathbf{g}_k; \nabla f(\mathbf{w}_k - \eta_k \mathbf{g}_k)\rangle \le \epsilon_k < \epsilon$. Using continuity again we also know $s_k < \eta_k$. Applying intermediate value theorem we obtain:

$$\langle \mathbf{g}_k; \nabla f(\mathbf{w}_k - s_k \mathbf{g}_k)\rangle = \epsilon.$$

Let $s = \inf s_k$. We distinguish two cases:

- $s = 0$, in which case (by passing to a further subsequence if necessary) we may assume $s_k \to 0$. Thus,

$$\begin{aligned}\epsilon = \langle \mathbf{g}_k; \nabla f(\mathbf{w}_k - s_k\mathbf{g}_k)\rangle &= \langle \mathbf{g}_k; \nabla f(\mathbf{w}_k)\rangle + \langle \mathbf{g}_k; \nabla f(\mathbf{w}_k - s_k\mathbf{g}_k) - \nabla f(\mathbf{w}_k)\rangle \\ &\geq \alpha\|\nabla f(\mathbf{w}_k)\|_\circ^2 - \delta_k - \|\mathbf{g}_k\| \cdot \|\nabla f(\mathbf{w}_k - s_k\mathbf{g}_k) - \nabla f(\mathbf{w}_k)\|_\circ \\ &\to \alpha\|\nabla f(\mathbf{w}_*)\|_\circ^2 > \epsilon,\end{aligned}$$

  contradiction.

- $s > 0$. Using mean value theorem there exists $\theta_k \in [0, 1]$ such that

$$f(\mathbf{w}_k - s\mathbf{g}_k) = f(\mathbf{w}_k) - s\,\langle \mathbf{g}_k; \nabla f(\mathbf{w}_k - \theta_k s\mathbf{g}_k)\rangle \leq f(\mathbf{w}_k) - \epsilon s,$$

  where the inequality follows from the minimality of s (see (6.3)). By definition $s \in [0, \eta_k]$, using the third condition we have

$$f(\mathbf{w}_k) - \epsilon s \geq f(\mathbf{w}_k - s\mathbf{g}_k) \geq f(\mathbf{w}_k - \eta_k\mathbf{g}_k) - \beta_k s \implies f(\mathbf{w}_{k+1}) \leq f(\mathbf{w}_k) - (\epsilon - \beta_k)s.$$

  As $\beta_k \to 0$, the above indicates $f(\mathbf{w}_k) \to -\infty$, contradicting to the assumption that $\mathbf{w}_k \to \mathbf{w}_*$.

The proof is now complete. ■

Most of the arguments here are due to Byrd and Tapia (1975) while Penot (2002) removed an implicit unnecessary assumption and introduced the relaxations. This result could be considered the final, definitive convergence result about metric gradient.

Needless to say, the following step size rules will satisfy the conditions in the theorem with $\epsilon_t = \beta_t = 0$:

- Cauchy's (global) rule: $\eta_t = \operatorname{argmin}_{\eta \geq 0} f(\mathbf{w}_t - \eta\mathbf{g}_t)$, i.e. η is a global minimizer of $f(\mathbf{w}_t - \eta\mathbf{g}_t)$.
- Curry's rule: $\eta_t = \inf\{\eta \geq 0 : \eta \text{ is a stationary point of } f(\mathbf{w}_t - \eta\mathbf{g}_t),\ \text{i.e. } \langle \mathbf{g}_t; \nabla f(\mathbf{w}_t - \eta\mathbf{g}_t)\rangle = 0\}$.
- Cauchy's local rule: $\eta_t = \inf\{\eta \geq 0 : \eta \text{ is a local minimizer of } f(\mathbf{w}_t - \eta\mathbf{g}_t)\}$.

Indeed, for large t, if $\nabla f(\mathbf{w}_t) \neq \mathbf{0}$, then we may assume w.l.o.g. that

$$\langle \mathbf{g}_t; \nabla f(\mathbf{w}_t)\rangle \geq \alpha\|\nabla f(\mathbf{w}_t)\| - \delta_t > 0,$$

for otherwise $\|\nabla f(\mathbf{w}_t)\| \leq \delta_t \to 0$ already. Using Taylor expansion and continuity we then know $\eta_t > 0$ in all cases since $f(\mathbf{w}_t - \eta\mathbf{g}_t)$ is strictly decreasing for small η.

Byrd, R. H. and R. A. Tapia (1975). "An extension of Curry's theorem to steepest descent in normed linear spaces". *Mathematical Programming*, vol. 9, pp. 247–254.
Penot, J.-P. (2002). "On the Convergence of Descent Algorithms". *Computational Optimization and Applications*, vol. 23, pp. 279–284.

Theorem 6.15: Convergence of metric gradient descent

In addition to the conditions in Theorem 6.14, if f is bounded from below on a set W that contains the iterates of Algorithm 6.8 and ∇f is bounded on W and uniformly continuous around W, then $\nabla f(\mathbf{w}_t) \to \mathbf{0}$.

Proof: Suppose to the contrary there exists an infinite subsequence $\mathbf{w}_k := \mathbf{w}_{t_k}$ such that $\|\nabla f(\mathbf{w}_k)\|_\circ \geq \delta > 0$. Let $\epsilon \in (0, \alpha\delta^2)$ and define s_k and $s = \inf s_k$ as in the proof of Theorem 6.14. Again, we distinguish two cases:

- $s = 0$, in which case we may assume $s_k \to 0$. Using the boundedness of $\mathbf{g}_k$ and the uniform continuity of ∇f around W we have for large k that

$$\epsilon = \langle \mathbf{g}_k; \nabla f(\mathbf{w}_k - s_k\mathbf{g}_k)\rangle > \langle \mathbf{g}_k; \nabla f(\mathbf{w}_k)\rangle - (\alpha\delta^2 - \epsilon) \geq \alpha\delta^2 - \delta_k - (\alpha\delta^2 - \epsilon) \to \epsilon.$$

- $s > 0$, in which case we have from the proof of Theorem 6.14 that

$$f(\mathbf{w}_{k+1}) \leq f(\mathbf{w}_k) - (\epsilon - \beta_k)s,$$

  contradicting the boundedness of f on W.

The proof is complete now. ■

A function f is uniformly continuous around a set W if for any $\epsilon > 0$ there exists $\delta > 0$ such that for all $\mathbf{w} \in W$ and $\|\mathbf{z} - \mathbf{w}\| \leq \delta$ we have $|f(\mathbf{w}) - f(\mathbf{z})| \leq \epsilon$.

Definition 6.16: Metric projection

Given an arbitrary norm and a closed set C, we define the associated metric projection:

$$\mathrm{P}_C(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{z} \in C} \|\mathbf{w} - \mathbf{z}\|.$$

The non-emptiness of the metric projection is guaranteed if the space is reflexive. However, we lose some key properties. For instance, the metric projection may no longer be nonexpansive even when the set C is convex.

Algorithm 6.17: Metric projected gradient descent (McCormick, 1969)

Algorithm: Metric projected gradient descent for constrained smooth minimization

Input: $\mathbf{w}_0 \in C$, norm $\|\cdot\|$

for $t = 0, 1, \ldots$ **do**

- $\mathbf{g}_t \leftarrow \overline{\nabla} f(\mathbf{w}_t)$ `// compute any metric gradient`
- $\eta_t \leftarrow \operatorname{argmin}_{\eta \geq 0} f(\mathrm{P}_C(\mathbf{w}_t - \eta \mathbf{g}_t))$ `// Cauchy's rule, can use Curry's sometimes`
- $\mathbf{w}_{t+1} \leftarrow \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathbf{g}_t)$ `// update`

McCormick, G. P. (1969). "Anti-Zig-Zagging by Bending". *Management Science*, vol. 15, no. 5, pp. 315–320.

Remark 6.18: Convergence of Algorithm 6.17

Convergence of Algorithm 6.17 with Cauchy's rule was established by Phelps (1985). Convergence under Curry's rule is significantly harder, and only partial results were obtained in McCormick and Tapia (1972) and Phelps (1986).

Phelps, R. R. (1985). "Metric Projections and the Gradient Projection Method in Banach Spaces". *SIAM Journal on Control and Optimization*, vol. 23, no. 6, pp. 973–977.
McCormick, G. P. and R. A. Tapia (1972). "The Gradient Projection Method under Mild Differentiability Conditions". *SIAM Journal on Control*, vol. 10, no. 1, pp. 93–98.
Phelps, R. P. (1986). "The Gradient Projection Method Using Curry's Steplength". *SIAM Journal on Control and Optimization*, vol. 24, no. 4, pp. 692–699.