Chapter IV
First-order methods

Skills to acquire

- Compute subdifferentials and proximals operators
- Proximal gradient algorithms
- Applications to well known optimization problems.

We consider a general convex optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \tag{IV.1}$$

This chapter presents first order algorithms, i.e., which make use of the function gradient. Such algorithms have low computational cost per iterations.

1 Gradient descent algorithms

1.1 Description

We suppose that f is differentiable with $\mathrm{dom}(f) = \mathbb{R}^n$. The gradient descent algorithm solves the problem (IV.1) using the core iteration k:

- **Choice of the step**: Choose $\tau^{(k)}$
- **Main iteration**: $\boxed{\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \tau^{(k)}\nabla f(\boldsymbol{x}^{(k)})}$

The basic algorithm have some variants which depend mainly on the choice of the step $\tau^{(k)}$. A simple *constant step*

$$\tau^{(k)} = \tau, \forall k,$$

can diverge if τ is too big and slow if τ is too small. On the other hand, an optimal *exact line search*

$$\tau^{(k)} = \min_{\tau} f(x^k - \tau \nabla f(x^l))$$

can be hard to compute, computationally demanding, or even does not exist.

Another popular approach is *Backtracking line search*, which consists in the following. At each iteration, start with $\tau^{(k)} = \tau_0$ (e.g., $\tau_0 = 1$). Then, take

$$\tau^{(k)} = \beta \tau^{(k)},$$

til it satifies the Armijo condition

$$f(x^{(k)} - t^{(k)} \nabla f(x^{(k)})) \leq \left(f(x^{(k)}) - \alpha t^{(k)} \|\nabla f(x)\|_2^2 \right),$$

where $0 < \beta < 1$ and $0 < \alpha \leqslant 1/2$ are some fixed parameters. One can simply use $\alpha = \beta = 0.5$. These simple τ updates tend to work well in practice.

1.2 Convergence

Gradient descent algorithms are inexpensive as they do not require second derivatives.

Theorem 1.1. *Suppose that f is convex and $\nabla f(x)$ is L-Lipschitz continuous*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leqslant L \|\boldsymbol{x} - \boldsymbol{y}\|_2, \quad \forall\, \boldsymbol{x}, \boldsymbol{y}.$$

If the optimal value is finite and attained at x^* then the gradient descent converges to x^*, with a rate $O(1/\epsilon)$[1] for constant step size and $O(\log(1/\epsilon))$ for backtracking line search.

For unconstrained problems, gradient descent is robust and less tuning. Thus, it is still widely used in practice. But, for a non-convex function f, the convergence is often slow and depends on the starting point and the scaling. In addition, it cannot handle nondifferentiable functions. There are many other better algorithms that deal with its shortcomings.

1.3 Accelerated gradient method

Heavy Ball Method In this technique, we add a momentum to gradient descent:

$$\boxed{x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}) + \beta_k (x^{(k)} - x^{(k-1)}),}$$

where momentum prevents oscillation.

Nesterov Extrapolation (A simple variant) Nesterov's exploration alternates between gradient updates and proper extrapolation

$$\begin{cases} x^{(k+1)} = y^{(k)} - \alpha_k \nabla f(y^{(k)}), \\ y^{(k+1)} = x^{(k+1)} + \frac{k}{k+3}(x^{(k+1)} - x^{(k)}). \end{cases}$$

Each iteration takes nearly same cost as GD but converge faster. Note that it is not a descent method.

[1] i.e., to get $f(x^{(k)}) - f(x^*) \leqslant \epsilon$, we need $O(1/\epsilon)$ iterations.

1.4 Quasi-Newton methods

By successive measurements of the gradient, Quasi-Newton methods build a quadratic model of the objective function that can be used to estimate the inverse of the Hessian. The algorithm is the following.

Algorithm IV.1: Quasi-Newton algorithm

```
/* Initialization                                                    */
```
Choose initial point $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ and a matrix $H_0 \succ 0$
```
/* Main loop                                                         */
```
for $k = 0, 1, 2, 3, \ldots, N$ **do**
```
    /* compute quasi-Newton direction                                */
```
    $\Delta x_k = -H_k^{-1} \nabla f x_k$
```
    /* determine step size (e.g., by backtracking line search)       */
```
    Compute t_k
```
    /* compute $x_{k+1}$                                             */
```
    $x_{k+1} = x_k + t_k + \Delta x_k$
```
    /* Update $H$                                                    */
```
    Compute H_{k+1}.

end

There are different update rules exist for H_{k+1}. One can also estimate H_k^{-1} or a factorization of H^k to simplify the computation of Δx_k. The BFGS method (Broyden-Fletcher-Goldfarb-Shanno) is the most popular quasi-Newton method.

BFGS update

$$H_{k+1} = H_{k+1} + \frac{yy^T}{y^T s} - \frac{H_k s s^T H_k}{s^T H_k s},$$

where

$$s = x_{k+1} - x_k, \quad y = \nabla f(x_{k+1}) - \nabla f(x_k).$$

Inverse update

$$H_{k+1}^{-1} = \left(I - \frac{sy^T}{y^T s}\right) H_k^{-1} \left(I - \frac{ys^T}{y^T s}\right) + \frac{ss^T}{y^T s}.$$

1.5 Nondifferentiable or constrained problems

For nondifferentiable or constrained problems, there are many methods that can be used, e.g., subgradient and proximal gradient methods.

2 Sub-gradient methods

Subgradients generalize gradients for optimizing convex functions that are not necessarily differentiable. In many situations, we deal with common operations that preserve convexity but not differentiability (e.g., the max-operation).

2.1 Sub-gradient

Recall that for differentiable f, that there is a single unique linear tangent $\nabla f(\boldsymbol{x})$ that under estimates the function (Figure IV.1):

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \quad \forall \boldsymbol{x}, \boldsymbol{y}.$$

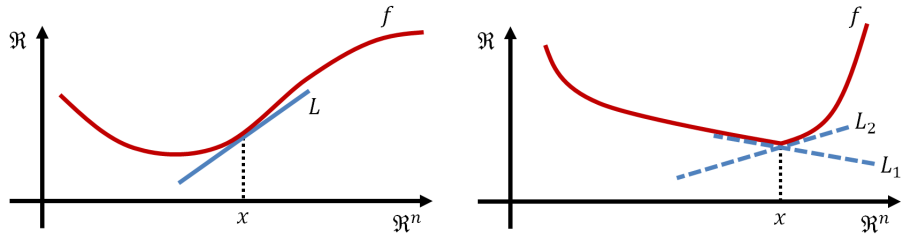


Fig. IV.1: Linear tangents under estimating a function.

Definition 2.1. *A* ***subgradient*** *of a function f at $\boldsymbol{x}$ is any $g \in \mathbb{R}^n$ such that*

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{y} - \boldsymbol{x}), \quad \forall\, \boldsymbol{y}$$

The set of all subgradients of convex f is called the ***subdifferential****:*

$$\partial f(\boldsymbol{x}) \overset{\text{def.}}{=} \left\{ \boldsymbol{g} \in \mathbb{R}^n \,;\, \forall\, \boldsymbol{y}, f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \boldsymbol{g}^T(\boldsymbol{y} - \boldsymbol{x}) \right\}.$$

> **Remark:** Informally, the "size" of $\partial f(x)$ controls how smooth f is at x. For nonconvex f, subgradients need not exist.

Example 2.1 (absolute value)**.** *For $f(x) = |x|$, it is easy to show that for $x \neq 0$, there is a unique subgradient $g = \operatorname{sign}(x)$. For $x = 0$, $\partial f(x) = [-1, 1]$.*

Example 2.2 (Subdifferential of norms at 0)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$. We can show that:*

$$\partial f(0) = \mathbb{B}_{\|.\|_*}(0, 1) = \left\{ \boldsymbol{g} \in \mathbb{R}^n \;\middle|\; \|g\|_* \leq 1 \right\},$$

where $\|.\|_*$ denotes the dual norm. In particular, for $g(x) = \|\boldsymbol{x}\|_1$, $\partial g(0) = [-1, 1]^n$.

Proposition 2.1 (Differentiability)**.** *f is differentiable at $x \iff \partial f(x) = \{\nabla f(x)\}$.*

Example 2.3. *Let* $f(x) = |\cdot|, x \in \mathbb{R}$.

$$\partial f(x) = \begin{cases} -1 & \text{if} \quad x < 0, \\ +1 & \text{if} \quad x > 0, \\ [-1,1] & \text{if} \quad x = 0. \end{cases}$$

Example 2.4 (Subdifferential of the ℓ_2-norm)**.** *Let* $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$.

$$\partial f(\boldsymbol{x}) = \begin{cases} \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2} & \boldsymbol{x} \neq 0, \\ \mathbb{B}_2(0,1) & \boldsymbol{x} = 0. \end{cases}$$

Theorem 2.2. *A subdifferential* $\partial f(x)$ *is closed and convex, even for nonconvex* f.

Theorem 2.3 (Existence)**.** *Le* f *be convex with a nonempty* $\mathrm{dom}(f)$*. Then,*

(a) $x \in \mathrm{relint}(\mathrm{dom}(f)) \implies \partial f(\boldsymbol{x}) \neq \emptyset$.
(b) $\partial f(\boldsymbol{x}) \neq \emptyset$ *and bounded* $\iff x \in \mathrm{int}(\mathrm{dom}(f))$.

Example 2.5.

$$f(x) = \begin{cases} -\sqrt{x}, x \geqslant 0 \\ +\infty, \ x < 0 \end{cases}; \quad \partial f(x) = \begin{cases} -\frac{1}{2\sqrt{x}}, x > 0 \\ \emptyset, \qquad x = 0 \end{cases}$$

Note that 0 *is not in the interior of* $\mathrm{dom}(f)$.

Theorem 2.4 (Directional Derivatives)**.** *Let* $\boldsymbol{x} \in \mathrm{int}(\mathrm{dom}(f))$*. Then, the directional derivative* $f'(\boldsymbol{x}; \boldsymbol{d})$ *is finite for all* $\boldsymbol{d} \in \mathbb{R}^n$*. In particular, we have*

$$f'(\boldsymbol{x}; \boldsymbol{d}) = \max_{s \in \partial f(x)} s^T \boldsymbol{d}$$

Example 2.6 (connection to convex geometry)**.** *For a non empty set* $C \subseteq \mathbb{R}^n$*,* $\partial \mathbb{1}_C(\boldsymbol{x}) = \mathcal{N}_C(\boldsymbol{x})$*. Recall that for* $\boldsymbol{x} \notin C, \partial \mathbb{1}_C(\boldsymbol{x}) = \mathcal{N}_C(\boldsymbol{x}) = \emptyset$ *by convention. As a special case,*

$$\partial \mathbb{1}_{\mathbb{B}(0,1)}(\boldsymbol{x}) = \mathcal{N}_{\mathbb{B}(0,1)}(\boldsymbol{x}) = \begin{cases} \left\{ \boldsymbol{y} \in \mathbb{R}^n \;\middle|\; \|\boldsymbol{y}\|_* \leq \boldsymbol{y}^T \boldsymbol{x} \right\}, & \|\boldsymbol{x}\| \leq 1 \\ \emptyset, & \|\boldsymbol{x}\| > 1. \end{cases}$$

2.2 Subdifferential calculus

In this section, we are present some practical rules for calculating subgradients. We can distinguish two types of rules giving weak (one subgradient only) or strong (all subgradients) results.

Rule 1: Conic combination

Theorem 2.5. *Let* $f, g : \mathbb{R}^n \to (-\infty, \infty]$ *be proper functions and let* $\alpha > 0$.

- *Multiplication by a Positive Scalar*

$$\forall\, \boldsymbol{x} \in \mathrm{dom}(f): \quad \partial(\alpha f(\boldsymbol{x})) = \alpha\, \partial f(\boldsymbol{x}).$$

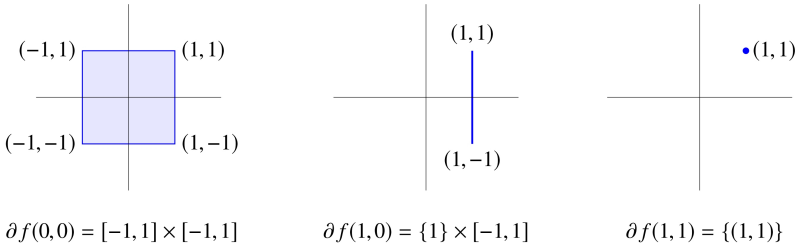
- *Summation*

$$\forall\, \boldsymbol{x} \in \mathrm{dom}(f) \cap \mathrm{dom}(g): \quad \partial f(\boldsymbol{x}) + \partial g(\boldsymbol{x}) \subseteq \partial(f+g)(\boldsymbol{x}),$$
$$\forall\, \boldsymbol{x} \in \mathrm{int}(\mathrm{dom}(f) \cap \mathrm{dom}(g)): \quad \partial(f+g)(\boldsymbol{x}) = \partial f(\boldsymbol{x}) + \partial g(\boldsymbol{x}).$$

The first summation rule is weak and the second is strong.

Example 2.7 (Subdifferentials of the ℓ_1-norm)**.** *Consider $f : \mathbb{R}^n \to \mathbb{R}$ given by $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$. Then, $f = \sum_{i=1}^n f_i$, where $f_i(\boldsymbol{x}) = |x_i|$. Hence,*

$$\partial f(\boldsymbol{x}) = \left\{ \boldsymbol{z} \in \mathbb{R}^n \;\middle|\; z_i = sign(x_i) \text{ if } x_i \neq 0,\ |z_j| \leq 1 \text{ if } x_j = 0 \right\}.$$



$\partial f(0,0) = [-1,1] \times [-1,1]$ $\partial f(1,0) = \{1\} \times [-1,1]$ $\partial f(1,1) = \{(1,1)\}$

Rule 2: Affine composition

Theorem 2.6. *Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be a proper function and $h(\boldsymbol{x}) = f(A\boldsymbol{x} + \boldsymbol{b})$. Assume h is proper, and let $\alpha > 0$.*

- *Weak result*

$$\forall\, \boldsymbol{x} \in \mathrm{dom}(f): \quad A^T(\partial f(A\boldsymbol{x} + \boldsymbol{b})) \subseteq \partial h(\boldsymbol{x}).$$

- *Strong result*

$$\forall\, \boldsymbol{x} \in \mathrm{int}(\mathrm{dom}(f)) \;/\; A(\boldsymbol{x}) + \boldsymbol{b} \in \mathrm{int}(\mathrm{dom}(f)): \quad \partial h(\boldsymbol{x}) = A^T(\partial f(A\boldsymbol{x} + \boldsymbol{b})).$$

Example 2.8. *Compute the subdifferential of $f(\boldsymbol{x}) = \|A\boldsymbol{x} + \boldsymbol{b}\|_1$.*

$$\partial f(\boldsymbol{x}) = \sum_{i \in I_1} \mathrm{sign}(\mathbf{a}_i^T \boldsymbol{x} + b_i)\mathbf{a}_i + \sum_{i \in I_2} [-\mathbf{a}_i, \mathbf{a}_i],$$

where $\mathbf{a}_i$ is the i^{th} column of A, $I_1 = \{i : \mathbf{a}_i^T \boldsymbol{x} + b_i \neq 0\}$, $I_2 = \{i : \mathbf{a}_i^T \boldsymbol{x} + b_i = 0\}$. A weak result indicating one possible subgradient is $A^T \mathrm{sign}(A\boldsymbol{x} + \boldsymbol{b}) \in \partial f(\boldsymbol{x})$.

Example 2.9. *Compute the subdifferential of $\|A\boldsymbol{x} + \boldsymbol{b}\|_2$.*

$$\partial f(\boldsymbol{x}) = \begin{cases} \frac{A^T(A\boldsymbol{x}+\boldsymbol{b})}{\|A\boldsymbol{x}+\boldsymbol{b}\|_2}, & A\boldsymbol{x} + \boldsymbol{b} \neq 0, \\ A^T \mathbb{B}_{\|.\|_2}(0,1), & A\boldsymbol{x} + \boldsymbol{b} = 0. \end{cases}$$

If a weak result is required, then $0 \in \partial f(\boldsymbol{x})$ for any $\boldsymbol{x}$ satisfying $A\boldsymbol{x} + \boldsymbol{b} = 0$.

Rule 3: Chain rule

Theorem 2.7 (Sub-differential calculus)**.** . *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and $g : \mathbb{R} \to \mathbb{R}$ be a nondecreasing convex function. Let $\boldsymbol{x} \in \mathbb{R}^n$, and suppose that g is differentiable at the point $f(\boldsymbol{x})$. Let $h = g \circ f$. Then*

$$\partial h(\boldsymbol{x}) = g'(f(\boldsymbol{x}))\partial f(\boldsymbol{x}).$$

Example 2.10 (subdifferential of $\|.\|_1^2$)**.** *Consider the function $h : \mathbb{R}^n \to \mathbb{R}$ given by $h(\boldsymbol{x}) = \|\boldsymbol{x}\|_1^2$, which can be written as the composition $h = g \circ f$, where $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ and $g(t) = \max\{t, 0\}^2$. Both f and g are real-valued convex functions, and g is nondecreasing and differentiable over $\mathbb{R}$ with derivative $g'(t) = 2\max\{t, 0\}$. Therefore, by the chain rule of subdifferential calculus, for any $x \in \mathbb{R}^n$,*

$$\partial h(\boldsymbol{x}) = g'(f(\boldsymbol{x}))\partial f(\boldsymbol{x}) = 2\max\{0, \|x\|_1\}\partial f(x),$$

$$\partial h(x) = 2\|x\|_1 \left\{z \in \mathbb{R}^n : z_i = sign(x_i), i \in I_1, |z_j| \leq 1, j \in I_2\right\},$$

where $I_1 = \{i : x_i \neq 0\}$ and $I_2 = \{i : x_i = 0\}$.

Example 2.11 (Distance to a convex set)**.** *Recall the distance function to a closed and convex set C:*

$$d_C(\boldsymbol{x}) = \min_{\boldsymbol{y} \in C} \ \|\boldsymbol{y} - \boldsymbol{x}\|_2 = \|\boldsymbol{x} - P_C(\boldsymbol{x})\|_2$$

where $P_C(\boldsymbol{x})$ is the projection of $\boldsymbol{x}$ onto C. We will show that

$$\partial d_C(\boldsymbol{x}) = \begin{cases} \left\{\frac{\boldsymbol{x} - P_C(\boldsymbol{x})}{\|\boldsymbol{x} - P_C(\boldsymbol{x})\|_2}\right\}, & \boldsymbol{x} \notin C, \\ \\ \mathcal{N}_C(\boldsymbol{x}) \cap \mathbb{B}(0, 1), & \boldsymbol{x} \in C. \end{cases}$$

We know that the function $\varphi_C(\boldsymbol{x}) = \frac{1}{2}d_C^2(\boldsymbol{x})$ is differentiable and

$$\forall\, \boldsymbol{x}, \quad \partial\varphi_C(\boldsymbol{x}) = \{\boldsymbol{x} - P_C(\boldsymbol{x})\}.$$

Note that $\varphi_C = g \circ d_C$, where $g(t) = \frac{1}{2}\max\{0, t\}^2$ is a nonincreasing real-valued convex differentiable function. Then by the chain rule:

$$\varphi_C(\boldsymbol{x}) = d_C(\boldsymbol{x})\ \partial d_C(\boldsymbol{x}).$$

If $\boldsymbol{x} \notin C$, then $d_C(\boldsymbol{x}) = 0$, and thus $\partial d_C(\boldsymbol{x}) = \left\{\frac{\boldsymbol{x} - P_C(\boldsymbol{x})}{d_C(\boldsymbol{x})}\right\}$. If $\boldsymbol{x} \in C$, we show the double inclusion.

Rule 4: Maximization

Theorem 2.8 (max rule of subdifferential calculus)**.** *Let $f_1, f_2, \ldots, f_m : \mathbb{R}^n \to (-\infty, \infty]$ be proper convex functions. Define $f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\}$. Then,*

$$\forall\, \boldsymbol{x} \in \bigcap_{i=1,..,m} \int \mathrm{dom}(f_i) : \quad \partial f(\boldsymbol{x}) = conv\left(\bigcup_{i \in I(\boldsymbol{x})} \partial f_i(\boldsymbol{x})\right),$$

where $I(\boldsymbol{x}) = \left\{i \;\middle|\; f_i(\boldsymbol{x}) = f(\boldsymbol{x})\right\}$, the "active" functions at $\boldsymbol{x}$. As a weak result; to compute a subgradient at $\boldsymbol{x}$, choose any $k \in I(\boldsymbol{x})$, any subgradient of f_k at $\boldsymbol{x}$.

Example 2.12 (max of two functions)**.** *Let $f_1, f_2 : \mathbb{R}^N \to \mathbb{R}$ convex, differentiable, and $f(\boldsymbol{x}) = \max\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x})\}$. Then,*

- *For $f_1(\boldsymbol{x}) > f_2(\boldsymbol{x})$, unique subgradient $g = \nabla f_1(\boldsymbol{x})$*
- *For $f_2(\boldsymbol{x}) > f_1(\boldsymbol{x})$, unique subgradient $g = \nabla f_2(\boldsymbol{x})$*
- *For $f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})$, $\partial f(\boldsymbol{x}) =$ the line segment joining $\nabla f_1(\boldsymbol{x})$ and $\nabla f_2(\boldsymbol{x})$.*

Example 2.13 (subdifferential of the max function)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\boldsymbol{x}) = \max\{x_1, x_2, \ldots, x_n\}$. Then we get:*

$$\partial f(\boldsymbol{x}) = \operatorname{conv}\left(\bigcup_{i \in I(\boldsymbol{x})} \partial f_i(\boldsymbol{x})\right) = \operatorname{conv}\left(\bigcup_{i \in I(\boldsymbol{x})} \{\boldsymbol{e}_i\}\right),$$

$$\partial f(\boldsymbol{x}) = \left\{ \sum_{i \in I(\boldsymbol{x})} \lambda_i \boldsymbol{e}_i : \sum_{i \in I(\boldsymbol{x})} \lambda_i = 1, \lambda_{\geq} 0 \right\}.$$

Example 2.14 (subdifferential of the ℓ_∞-norm)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty$.*

$$\partial f(\boldsymbol{x}) = \begin{cases} \mathbb{B}_{\|\cdot\|_1}(0,1), & \boldsymbol{x} = 0, \\ \left\{ \sum_{i \in I(\boldsymbol{x})} \lambda_i \operatorname{sign}(x_i) \boldsymbol{e}_i : \sum_{i \in I(\boldsymbol{x})} \lambda_i = 1, \lambda_j \geq 0, j \in I(\boldsymbol{x}) \right\}, & \boldsymbol{x} \neq 0. \end{cases}$$

Example 2.15. ***piecewise-linear function***

$$f(\boldsymbol{x}) = \max_{i=1,\ldots,m} a_i^T \boldsymbol{x} + b$$

the subdifferential at $\boldsymbol{x}$ is a polyhedron

$$\partial f(\boldsymbol{x}) = \operatorname{conv}\{a_i | i \in I(\boldsymbol{x})\}$$

Theorem 2.9 (weak maximum rule of subdifferential calculus)**.** *Let I be an arbitrary set, and suppose that any $i \in I$ is associated with a proper convex function $f_i : \mathbb{R}^n \to (-\infty, \infty]$. Let $f(\boldsymbol{x}) = \max_{i \in I} f_i(\boldsymbol{x})$. Then for any $\boldsymbol{x} \in \operatorname{dom}(f)$:*

$$\operatorname{conv}\left(\bigcup_{i \in I(\boldsymbol{x})} \partial f_i(\boldsymbol{x})\right) \subseteq \partial f(\boldsymbol{x}),$$

where $I(\boldsymbol{x}) = \left\{ i \in I \;\middle|\; f(\boldsymbol{x}) = f_i(\boldsymbol{x}) \right\}$. Usually we get equality, but requires some technical conditions to hold.

Rule 5: Minimization

Theorem 2.10 (Sub-differential calculus)**.** $f(x) = \inf_y h(x,y)$, h *convex. To find a subgradient at* $\hat{x}$ *(weak result), find* $\hat{y}$ *that minimizes* $h(\hat{x},y)$ *(assuming minimum is attained) and then find subgradient* $(g,0) \in \partial h(\hat{x},\hat{y})$.

Example 2.16. *Find a subgradient of*

$$f(x) = \inf_{y\in\mathcal{C}} \|x-y\|_2$$

where $\mathcal{C}$ *is a closed convex set.*

- *if* $f(\hat{x}) = 0$ *(that is,* $\hat{x} \in \mathcal{C}$*), take* $g = 0$.
- *if* $f(\hat{x}) > 0$*, find projection* $\hat{y} = P(\hat{x})$ *on* $\mathcal{C}$ *and take*

$$g = \frac{\hat{x}-\hat{y}}{\|\hat{x}-\hat{y}\|_2}$$

2.3 Optimality conditions

Subdifferential sets are extremely useful in characterizing minima points. The following theorem is simple but extremely important.

Theorem 2.11 (Fermat's optimality condition)**.** *Let* $f : \mathbb{R}^n \to (-\infty,\infty]$ *be a proper convex function. Then,*

$$\boldsymbol{x}^\star \in \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \iff 0 \in \partial f(\boldsymbol{x}^\star).$$

Proof. One has

$$\boldsymbol{x}^\star \in \operatorname{argmin} f \quad \Leftrightarrow \quad \Big(\forall\, \boldsymbol{y}, f(\boldsymbol{x}^\star) \leqslant f(\boldsymbol{y}) + \mathbf{0}^T(\boldsymbol{x}^\star - \boldsymbol{y})\Big) \quad \Leftrightarrow \quad 0 \in \partial f(\boldsymbol{x}^\star).$$

□

Example 2.17 (medians)**.** *Suppose that we are given* n *different 20 and ordered numbers* $a_1 < a_2 < \cdots < a_n$*. Denote* $A = \{a_1, a_2, \ldots, a_n\} \subset \mathbb{R}$*. The set of possible medians is the optimal solution set of the problem*

$$\min_{x\in\mathbb{R}} f(x),$$

where $f(x) = \sum_{i=1}^{n} |x - a_i| = f_i(x)$.

$$\partial f(x) = \begin{cases} 2i-n, & x \in (a_i, a_{i+}), \\ 2i-1-n+[-1,1], & x = a_i, \\ -n, & x < a_1, \\ n, & x > a_n. \end{cases}$$

$0 \in \partial f(a_i)$ *if and only if* $|2i-1-n| \leqslant 1$*, which is equivalent to* $\frac{n}{2} \leqslant i \leqslant \frac{n}{2}+1$ *and* $0 \in \partial f(x)$ *for some* $x \in (a_i, a_{i+1})$ *if and only if* $i = \frac{n}{2}$*. We can thus conclude that if* n *is odd, then the only optimal point is* $\frac{a_{n+1}}{2}$ *and when* n *is even, the optimal set is the interval* $[a_{\frac{n}{2}}, a_{\frac{n}{2}+1}]$*, establishing the fact that the optimal set is exactly the set of medians.*

Theorem 2.12 (optimality conditions for composite problems)**.** *Let* $f, g : \mathbb{R}^n \to (-\infty, \infty]$ *be proper functions such that* $\mathrm{dom}(g) \subset \mathrm{int}(\mathrm{dom}(f))$*, g is convex. Consider the problem*

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) + g(\boldsymbol{x})$$

If $x^* \in dom(g)$ is a local optimal solution for which f is differentiable, then

$$-\nabla f(\boldsymbol{x}^*) \in \partial g(\boldsymbol{x}^*).$$

Such points are called ***stationary points****. This condition is sufficient when f is convex. In addition, the stationary points are global optimal solutions.*

Example 2.18 (Convex constraints)**.** *For f convex and differentiable and convex set C:*

$$x^* \in \operatorname*{argmin}_{x \in C} \ f(x) \iff \exists g \in \partial f(x^*) / -g \in \mathcal{N}_C(x^*).$$

Example 2.19 (Lasso)**.** *Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$:*

$$\operatorname*{argmin}_{\beta} \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$

where $\lambda \geqslant 0$. Subgradient optimality:

$$0 \in \partial\left(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right) \iff 0 \in -X^T(y - X\beta) + \lambda\partial\|\beta\|_1$$

$$0 \in \partial\left(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right) \iff \exists v \in \partial\|\beta\|_1 : \ X^T(y - X\beta) = \lambda v,$$

$$v_i = \begin{cases} \{1\} & \text{if } \beta_i > 0 \\ \{-1\} & \text{if } \beta_i < 0, \\ = [-1, 1] & \text{if } \beta_i = 0 \end{cases} \quad i = 1, ..., p$$

Write X_1, ..., X_p for columns of X. Then our condition reads:

$$\begin{cases} X_i^T(y - X\beta) & = \lambda \ \mathrm{sign}(\beta_i) \ \text{if } \beta_i \neq 0 \\ |X_i^T(y - X\beta)| & \leqslant \lambda \qquad\qquad\quad \text{if } \beta_i = 0 \end{cases}$$

Note: subgradient optimality conditions don't lead to closed-form expression for a lasso solution. However, they do provide a way to check lasso optimality.

Example 2.20 (Soft-thresholding)**.** *Simplified lasso problem with $X = I$:*

$$\operatorname*{argmin}_{\beta} \ \frac{1}{2}\|y - \beta\|_2^2 + \lambda\|\beta\|_1,$$

$$\beta = S_\lambda(y); \quad [S_\lambda(y)]_i = \begin{cases} y_i - \lambda \ \text{if } y_i > 0 \\ 0 \qquad \text{if } -\lambda \leqslant y_i < \lambda, \\ y_i + \lambda \ \text{if } y_i < -\lambda \end{cases} \quad i = 1, ..., n$$

Example 2.21. *Consider the problem*

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad f(x) + \lambda \|x\|_1 \tag{IV.2}$$

where $f : \mathbb{R}^n \to (-\infty, \infty]$ *is an extended real-valued function. A point* $x^* \in int(dom(f))$ *in which* f *is differentiable is a stationary point if* $- \bigtriangledown f(x^*) \in \lambda \partial g(x^*)$, *where* $g(.) = \|.\|_1$. *So, we obtain that* x^* *is a stationary point if*

$$\frac{\partial f(x^*)}{\partial x_i} \begin{cases} = -\lambda & if \ \ x_i^* > 0 \\ = \lambda & if \ \ x_i^* < 0 \\ \in [-\lambda, \lambda] & if \ \ x_i^* = 0 \end{cases}$$

This is a necessary condition for x^* *to be a local minimum. If* f *is also convex, then it is a necessary and sufficient condition for* x^* *to be a global optimal solution.*

2.4 Sub-gradient descent

When f is non differentiable, one can use a sub-gradient instead of the gradient to define a minimization.

Algorithm IV.2: Sub-gradient descent

```
/* Main loop                                                    */
for k = 0, 1, 2, 3, ..., N do
    /* Main iteration                                           */
    x^(k+1) = x^(k) − τ^(k) g^(ℓ)   where   g^(ℓ) ∈ ∂f(x^(ℓ)).
    /* Stopping criterion                                       */
    if ‖x^(k+1) − x^(k+1)‖_2 < ϵ then
        Stop
```

But, subgradients are not necessarily descent directions asshown on the example below. For instance, for $f(\boldsymbol{x}) = |x_1| + 3\,|x_2|$ at $\boldsymbol{x} = (1, 0)$, we have:

- $\mathbf{g}_1 = (1, 0) \in \partial f(x)$, and $-\mathbf{g}_1$ is a descent direction
- $\mathbf{g}_2 = (1, 3) \in \partial f(x)$, but $-\mathbf{g}_g$ is not a descent direction

So we keep track of best iterate $x_{best}^{(k)}$ among $x^{(0)}$, ..., $x^{(k)}$ so far, i.e., $f(x_{best}^{(k)}) = \min\limits_{i=0,\ldots,k} f(x^{(i)})$.

Step size rules

- fixed step: τ_k constant
- fixed length: $\tau^{(k)} g^{(\ell)} = \|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k+1)}\|_2$ is constant
- diminishing: $\tau_k \to 0$ and $\sum\limits_{k=0}^{\infty} \tau_k = 0$.

Theorem 2.13. *If* $\sum_\ell \tau_\ell = +\infty$ *and* $\sum_\ell \tau_\ell^2 < +\infty$, *then* $x^{(\ell)}$ *converges to a minimizer of* f.

Analysis The subgradient method is not a descent method but the distance to the optimal goes to zero. As for the gradient method, the convergence is ensured when f is L-Lipschitz near the optimum. But, in general, this method performs poorly and convergence can be very slow. However, it handles general nondifferentiable convex problem, especially when ∂f is easy to compute, and often leads to very simple algorithms. Another option could be trying a splitting strategy as we will show later.

3 Proximal Algorithms

Proximal mapping are fundamental tools in convex optimization. Let's start with a proximal view of gradient descent.

3.1 Proximal Mapping

Consider the minimization problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}).$$

We perform quadratic approximation, replacing usual Hessian $\nabla^2 f(x)$ by $\frac{1}{t}I$,

$$f(y) \approx f(x) + \nabla f(x)^T(y-x) + \frac{1}{2t}\|y-x\|.$$

This can be seen as a linear approximation of f for which we add a proximity term to x, with weight $1/(2t)$. So, the update rule can be written as

$$\boldsymbol{x}^{(k+1)} = \operatorname*{argmin}_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}^{(k)}) + \nabla f(\boldsymbol{x}^{(k)})^T(\boldsymbol{x}-\boldsymbol{x}^{(k)}) + \frac{1}{2t^{(k)}}\|\boldsymbol{x}-\boldsymbol{x}^{(k)}\|_2^2 \right\}.$$

Definition 3.2 (proximal mapping)**.** *Given a function $f : \mathbb{R}^n \to (-\infty, \infty]$, the proximal mapping of f is the operator given by*

$$\operatorname{prox}_f(\boldsymbol{x}) \overset{\text{def.}}{=} \operatorname*{argmin}_{\boldsymbol{z}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{z}\|_2^2 + f(\boldsymbol{z}). \tag{IV.3}$$

Example 3.1. $f(x) = \lambda|x|$. $\operatorname{prox}_f(x) = \begin{cases} x-\lambda, x > \lambda \\ x+\lambda, x < -\lambda \\ 0, \quad |x| \le \lambda. \end{cases}$ *, i.e.,* $\operatorname{prox}_f = S_\lambda$ *the soft-thresholding operator.*

Example 3.2. *Consider the following functions from $\mathbb{R}$ to $\mathbb{R}$,$\lambda > 0$ and $\mu \in \mathbb{R}$,*

$$g_1(x) = 0;\ g_2(x) = \begin{cases} 0, & x \neq 0 \\ -\lambda, & x = 0 \end{cases};\ g_3(x) = \begin{cases} 0, & x \neq 0 \\ \lambda, & x = 0 \end{cases};\ g_4(x) = \begin{cases} \mu x, & x \geq 0, \\ \infty, & x < 0, \end{cases};$$

$$\operatorname{prox}_{g_1}(x) = \{x\}; \quad \operatorname{prox}_{g_2}(x) = \begin{cases} \{0\}, & |x| < \sqrt{2\lambda} \\ \{x\}, & |x| > \sqrt{2\lambda}, \\ \{0, x\}, & |x| = \sqrt{2\lambda}, \end{cases}; \quad \operatorname{prox}_{g_3}(x) = \begin{cases} \{x\}, & x \neq 0 \\ \emptyset, & x = 0 \end{cases}$$

$\operatorname{prox}_{g_4}(x) = 0.$

Example 3.3 (Convex Quadratic). *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be given by* $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c$, *where* $A \in \mathbb{S}^n_+$, $\boldsymbol{b} \in \mathbb{R}^n$, *and* $c \in \mathbb{R}$. *Then,*

$$\mathrm{prox}_f(\boldsymbol{x}) = (A + I)^{-1}(\boldsymbol{x} - \boldsymbol{b}).$$

Theorem 3.14. *If* f *is proper closed and convex, then* $\mathrm{prox}_f(x)$ *is a singleton. It can be evaluated efficiently for many widely used functions (in particular, regularizers).*

3.2 Prox Calculus Rules

Rule 1: Separable functions

Theorem 3.15. *Suppose that* $f : \mathbb{R}^{n_1} \times \ldots \times \mathbb{R}^{n_m} \to (-\infty, \infty]$ *is given by*

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \sum_{i=1}^{m} f_i(\boldsymbol{x}_i), \quad x_i \in \mathbb{R}^{n_i}$$

Then for any $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$,

$$\mathrm{prox}_f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \mathrm{prox}_{f_1}(\boldsymbol{x}_1) \times \cdots \times \mathrm{prox}_{f_m}(\boldsymbol{x}_m).$$

Example 3.4 (ℓ_1-norm). *Suppose that* $g : \mathbb{R}^n \to \mathbb{R}$ *is given by* $g(x) = \lambda\|\boldsymbol{x}\|_1$, *where* $\lambda > 0$. *Then,* $g(\boldsymbol{x}) = \sum_{i=1}^{n} \varphi(x_i)$, *and* prox_g *is the soft thresholding.*

Rule 2: Scaling and translation

Theorem 3.16. *Let* $g : \mathbb{R}^n \to (-\infty, \infty]$ *be a proper function,* $\lambda \neq 0$, *and* $\mathbf{a} \in \mathbb{R}^n$. *Define* $f(\boldsymbol{x}) = g(\lambda\boldsymbol{x} + \mathbf{a})$ *and* $h(\boldsymbol{x}) = \lambda g(\boldsymbol{x}/\lambda)$. *Then*

$$\mathrm{prox}_f(\boldsymbol{x}) = \frac{1}{\lambda}(\mathrm{prox}_{\lambda^2 g}(\lambda\boldsymbol{x} + \mathbf{a}) - \mathbf{a}),$$

$$\mathrm{prox}_h(\boldsymbol{x}) = \lambda(\mathrm{prox}_{g/\lambda}(\boldsymbol{x}/\lambda).$$

Example 3.5. *Consider the function* $f : \mathbb{R}^n \to (-\infty, \infty]$ *given for any*

$$f(x) = \begin{cases} \mu x, 0 \leq x \leq \alpha, \\ \\ \infty \quad else, \end{cases}$$

where $\mu \in \mathbb{R}$ *and* $\alpha \geq 0$.

$$\mathrm{prox}_f(x) = \min\{\max\{x - \mu, 0\}, \alpha\}.$$

Rule 3: Norm composition

Theorem 3.17. *Let* $f : \mathbb{R}^m \to (-\infty, \infty]$ *be given by* $f(\boldsymbol{x}) = g(\|\boldsymbol{x}\|)$ *where* $g : \mathbb{R}^m \to (-\infty, \infty]$ *is a proper closed convex function satisfying* $\mathrm{dom}(g) \subseteq [0, \infty)$. *Then*

$$\mathrm{prox}_f(\boldsymbol{x}) = \begin{cases} \mathrm{prox}_g(\|\boldsymbol{x}\|)\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, & \boldsymbol{x} \neq 0, \\ \\ \{u \in \mathbb{R}^n : \|u\| = \mathrm{prox}_g(0)\}, & \boldsymbol{x} = 0. \end{cases}$$

Example 3.6 (prox of Euclidean norm). *Let* $f(\boldsymbol{x}) = \lambda\|x\|$.

$$\mathrm{prox}_f(\boldsymbol{x}) = \left(1 - \frac{\lambda}{\max\{\|x\|, \lambda\}}\ \boldsymbol{x}\right)$$

Rule 4: Quadratic addition

Theorem 3.18 (Quadratic addition)**.** *If* $f(x) = g(x) + \frac{\rho}{2}\|x - a\|_2^2$*, then*

$$\mathrm{prox}_f(x) = \mathrm{prox}_{\frac{1}{1+\rho}g}\left(\frac{1}{1+\rho}\boldsymbol{x} + \frac{\rho}{1+\rho}\mathbf{a}\right)$$

Theorem 3.19. *Properties*

- ***Contraction****:*
$$\|\mathrm{prox}_f(x) - \mathrm{prox}_f(y)\| \leqslant \|x - y\|.$$
- ***Subgradient characterization****:*
$$u = \mathrm{prox}_f(x) \iff x - u \in \partial f(u)$$
- ***Interpretation of prox via resolvant of subdifferential operator****:*
$$z = \mathrm{prox}_f(x) \iff z = \underbrace{(\mathrm{Id} + \partial f^{-1})}_{\textit{Resolvent of operator of } \partial f}(x)$$

3.3 Algorithms

3.3.1 Proximal Point Algorithm

One has the following equivalence

$$\begin{aligned} x^\star \in \operatorname{argmin} f \quad &\Leftrightarrow \quad 0 \in \partial f(x^\star) \quad \Leftrightarrow \quad x^\star \in (\mathrm{Id} + \tau\partial f)(x^\star) \\ &\Leftrightarrow \quad x^\star = (\mathrm{Id} + \tau\partial f)^{-1}(x^\star) = \mathrm{Prox}_{\tau f}(x^\star). \end{aligned}$$

This suggest the following *proximal point algorithm*.

$$\boxed{x^{(k+1)} = \mathrm{Prox}_{\tau_\ell f}(x^{(k)}).}$$

On contrast to the gradient descent fixed point scheme, the proximal point method is converging for any sequence of steps.

Theorem 3.20. *If* $0 < \tau_{\min} \leqslant \tau_\ell \leqslant \gamma_{\max} < +\infty$*, then* $x^{(\ell)} \to x^\star$ *a minimizer of* f*.*

3.3.2 Proximal gradient descent (Forward-Backward)

It is not always possible to compute the proximal operator of the objective function. We will consider here a class of problems by imposing some structure on the function to be minimized. We consider functions F of the form

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}),$$

where f is convex differentiable and g is convex, not necessarily differentiable.

$$\begin{aligned} x^\star \in \operatorname{argmin} f+g \quad &\Leftrightarrow \quad 0 \in \nabla f(x^\star) + \partial g(x^\star) \\ &\Leftrightarrow \quad x^\star - \tau \nabla f(x^\star) \in (\mathrm{Id} + \tau \partial g)(x^\star) \\ &\Leftrightarrow \quad x^\star = (\mathrm{Id} + \tau \partial g)^{-1} \circ (\mathrm{Id} - \tau \nabla f)(x^\star). \end{aligned}$$

This fixed point suggests the following *Forward-Backward algorithm.*

$$\boxed{\mathbf{x}^{(k+1)} = \operatorname{prox}_{t_k g}(\mathbf{x}^{(k)} - t_k \nabla f(\mathbf{x}^{(k)})).}$$

As for the gradient descent, we can apply a Nesterov acceleration to define the *Accelerated proximal gradient method*

$$\boxed{v = x^{(k-1)} + \frac{k-2}{k+1}\left(x^{(k-1)} - x^{(k-2)}\right)}$$
$$\boxed{x^{(k)} = \operatorname{Prox}_{t_k}\left(v - t_k \nabla g(x)\right)}$$

Example 3.7 (Lasso)**.** *Given $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$:*

$$\min_\beta \ \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{f(\beta)} + \underbrace{\lambda \|\beta\|_1}_{g(\beta)},$$

$$\operatorname{prox}_{tg}(x) = \operatorname*{argmin}_z \ \frac{1}{2t}\|\beta - z\|_2^2 + \lambda \|z\|_1 = S_{\lambda t}(\beta)$$

Hence proximal gradient update is:

$$\beta^+ = S_{\lambda t}\left(\beta + tX^T(y - X\beta)\right) \tag{IV.4}$$

Often called the ***iterative soft-thresholding algorithm*** *(**ISTA**).*
*Applying acceleration gives us **FISTA** (F is for Fast): (Beck and Teboulle '09)*

$$v = \beta^{(k-1)} + \frac{k-2}{k+1}\left(\beta^{(k-1)} - \beta^{(k-2)}\right)$$

$$\beta^{(k)} = \operatorname{prox}_{\lambda t_k}\left(v + t_k X^T(y - Xv)\right), \quad k = 1, 2, 3, \ldots$$

or

$$x^{(k+1)} = \operatorname{prox}_{\eta_k h}(y^k - \eta_k \nabla f(y^k))$$
$$y^{(k+1)} = x^{(k+1)} + \frac{\theta_k - 1}{\theta_{k+1}}(x^{(k+1)} - x^{(k)})$$
$$y^0 = x^0, \theta_0 = 1, \theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2}$$

Example 3.8 (Projected gradient descent)**.** *Given closed, convex set $C \in \mathbb{R}$,*

$$\min_{x \in C} g(x) \iff \min_x g(x) + \mathbb{1}_C(x) \tag{IV.5}$$

$$x^+ = P_C\left(x - t\nabla g(x)\right) \tag{IV.6}$$

3.3.3 Douglas-Rachford

We reconsider here the structured minimization problem

$$\underset{\boldsymbol{x} \in \mathbb{R}^n}{\text{minimize}} \quad F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}), \tag{IV.7}$$

where f, g are closed convex, but on contrary to the Forward-Backward, no smoothness is imposed on f. We here suppose that we can compute easily the proximal map of f and g.

Algorithm IV.3: Douglas-Rachford method

Data: $\min_x \; f(\mathbf{x}) + g(\mathbf{x})$

`/* Main loop */`

for $k = 0, 1, 2, 3, \ldots, N$ **do**

`/* Douglas-Rachford iterations */`

$\boxed{u^{(k+1)} = \text{prox}_g(x^{(k)} + w^{(k)})}$

$\boxed{x^{(k+1)} = \text{prox}_f(u^{(k+1)} - w^{(k)})}$

$\boxed{w^{(k+1)} = w^{(k)} + (x^{(k+1)} - u^{(k+1)})}$

`/* Stopping criterion */`

if $\|x^{(k+1)} - x^{(k+1)}\|_2 < \boldsymbol{\varepsilon}$ **then**

Stop

Note that it is of course possible to inter-change the roles of f and g, which defines another set of iterations.