An introduction to nonsmooth optimisation

Clarice Poon
University of Bath

March 13, 2020

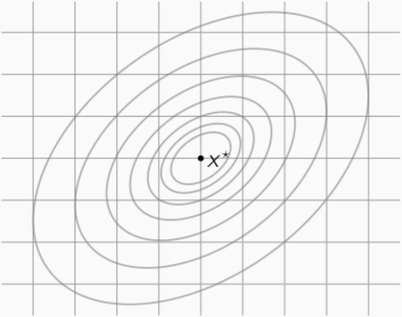
Outline

Unconstrained smooth optimisation

We first consider minimising

$$\min_{x \in \mathbb{R}^n} F(x)$$

where $F : \mathbb{R}^n \to \mathbb{R}$ is a proper, convex and differentiable function.



Unconstrained smooth optimisation

We first consider minimising

$$\min_{x\in\mathbb{R}^n} F(x)$$

where $F : \mathbb{R}^n \to \mathbb{R}$ is a proper, convex and differentiable function.

- Assume that the set of minimisers is nonempty,

$$\operatorname{argmin}(F) = \left\{ x \in \mathbb{R}^n \; ; \; F(x) = \min_{x\in\mathbb{R}^n} F(x) \right\} \neq \emptyset$$

- $x_* \in \operatorname{argmin}(F)$ has no closed form expression in general.
- We will consider iterative algorithms which start at a point x_0 and build a sequence x_k which converge to a minimiser.

Descent methods for smooth problems

General iterative method

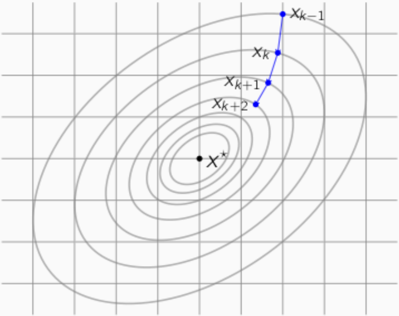
Initialization $x_0 \in \mathrm{dom}(F)$

while *stopping criterion not satisfied* **do**

  choose step-size $\gamma_k > 0$ and a search direction d_k

  Update $x_{k+1} = x_k - \gamma_k d_k$

end

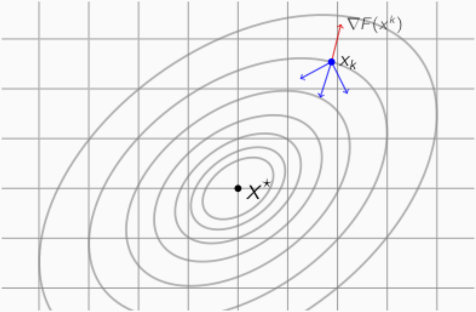


Choice of d_k to enforce descent

Descent methods

An algorithm is called a descent method if $F(x_{k+1}) < F(x_k)$.

- $\varphi_k(\gamma) \overset{\text{def.}}{=} F(x_k + \gamma d_k)$ is a decreasing function
- So, for d_k to be a descent direction, we need $\varphi_k'(0) = \langle \nabla F(x_k), d_k \rangle < 0$.



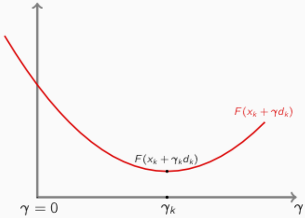
Step-size (Choice of γ_k)

1. Fixed step size $\gamma_k \equiv \gamma$.

Step-size (Choice of γ_k)

1. Fixed step size $\gamma_k \equiv \gamma$.
2. Line search along the direction d_k, i.e. minimise

$$\gamma_k = \text{argmin}_\gamma\, \varphi_k(\gamma) \stackrel{\text{def.}}{=} F(x_k + \gamma d_k)$$



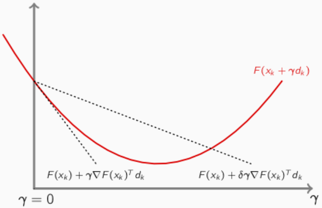
Step-size (Choice of γ_k)

1. Fixed step size $\gamma_k \equiv \gamma$.
2. Line search along the direction d_k, i.e. minimise

$$\gamma_k = \operatorname{argmin}_\gamma \varphi_k(\gamma) \stackrel{\text{def.}}{=} F(x_k + \gamma d_k)$$

3. Backtracking: Given direction d_k, choose $\delta \in (0,1)$ and $\beta \in (0,1)$ and let $\gamma = 1$.

$$\text{while } F(x_k + \gamma d_k) > F(x_k) + \delta\gamma\langle \nabla F(x_k),\, d_k\rangle : \ \gamma = \beta\gamma.$$



Stopping criterion

For some tolerance level $\varepsilon > 0$, we could consider:

Stopping criterion

For some tolerance level $\varepsilon > 0$, we could consider:

- Function value: $F(x_{k+1}) - F(x_k) \leqslant \varepsilon$.

Stopping criterion

For some tolerance level $\varepsilon > 0$, we could consider:

- Function value: $F(x_{k+1}) - F(x_k) \leqslant \varepsilon$.
- Sequence residue: $\|x_k - x_{k+1}\| \leqslant \varepsilon$.

Stopping criterion

For some tolerance level $\varepsilon > 0$, we could consider:

- Function value: $F(x_{k+1}) - F(x_k) \leqslant \varepsilon$.
- Sequence residue: $\|x_k - x_{k+1}\| \leqslant \varepsilon$.
- Optimality condition: $\|\nabla F(x_k)\| \leqslant \varepsilon$.

Outline

Gradient descent

The method of steepest descent or gradient descent chooses $d_k = -\nabla F(x_k)$. If x_k is not a stationary point, then $\nabla F(x_k) \neq 0$ and

$$\langle \nabla F(x_k),\, d_k \rangle = -\left\| \nabla F(x_k) \right\|^2 < 0$$

Gradient descent

Initialization $x_0 \in \mathrm{dom}(F)$

while *stopping criterion not satisfied* **do**

  choose step-size $\gamma_k > 0$

  Update $x_{k+1} = x_k - \gamma_k \nabla F(x_k)$

end

NB: method may not converge if γ_k is too large.

Convergence of gradient descent

Let $F \in \mathcal{C}^1$, ∇F is L-Lipschitz and $\operatorname{argmin}(F) \neq \emptyset$. Let $\gamma_k \equiv \gamma$.

General case

Choosing fixed $\gamma \in (0, 2/L)$, $\nabla F(x_k) \to 0$.

Convergence of gradient descent

Let $F \in \mathcal{C}^1$, ∇F is L-Lipschitz and $\operatorname{argmin}(F) \neq \emptyset$. Let $\gamma_k \equiv \gamma$.

General case

Choosing fixed $\gamma \in (0, 2/L)$, $\nabla F(x_k) \to 0$.

Convex case

Suppose that F is convex and let x_* be a minimiser. For $\gamma \in (0, 2/L)$,

$$F(x_k) - F(x_*) \leqslant \frac{\|x_0 - x_*\|^2}{\theta(k+1)}, \quad \text{where} \quad \theta = \gamma(1 - \gamma L/2). \tag{2.1}$$

Convergence of gradient descent

Let $F \in \mathcal{C}^1$, ∇F is L-Lipschitz and $\operatorname{argmin}(F) \neq \emptyset$. Let $\gamma_k \equiv \gamma$.

General case

Choosing fixed $\gamma \in (0, 2/L)$, $\nabla F(x_k) \to 0$.

Convex case

Suppose that F is convex and let x_* be a minimiser. For $\gamma \in (0, 2/L)$,

$$F(x_k) - F(x_*) \leqslant \frac{\|x_0 - x_*\|^2}{\theta(k+1)}, \quad \text{where} \quad \theta = \gamma(1 - \gamma L/2). \tag{2.1}$$

Strongly-convex case

If F is α-strongly convex, that is, $F(x) - \alpha \|x\|^2/2$ is convex (if $F \in \mathcal{C}^2$, equivalent to $\nabla^2 F(x) \geqslant \alpha \mathrm{Id}$), then

$$\|x_{k+1} - x_*\| \leqslant \max(1 - \gamma\alpha, \gamma L - 1)\|x_k - x_*\|.$$

Best constant is $\gamma = 2/(L + \alpha)$:

$$\|x_k - x_*\| \leqslant q^k \|x_0 - x_*\| \quad \text{where} \quad q = (L - \alpha)/(L + \alpha) \in (0, 1).$$

Suppose that x_k is an element of

$$x_0 + \mathrm{Span}\{\nabla F(x_0), \nabla F(x_1), \ldots, \nabla F(x_{k-1})\}. \tag{2.2}$$

Suppose that x_k is an element of

$$x_0 + \mathrm{Span}\{\nabla F(x_0), \nabla F(x_1), \ldots, \nabla F(x_{k-1})\}. \tag{2.2}$$

Theorem 2.1 (Nesterov's lower complexity bound)

For any $n \geqslant 2$ and $x_0 \in \mathbb{R}^n$, $L > 0$ and $k < n$, there exists a convex one-time continuously differentiable function F with L-Lipschitz continuous gradient, such that for any algorithm satisfying (2.2), we have

$$F(x_k) - F(x_*) \geqslant \frac{L\left\|x_0 - x_*\right\|^2}{8(k+1)^2}$$

where x_* denotes a minimiser of F.

- This result is valid only when the number of iterations is smaller than the problem size.

Suppose that x_k is an element of

$$x_0 + \mathrm{Span}\{\nabla F(x_0), \nabla F(x_1), \ldots, \nabla F(x_{k-1})\}. \tag{2.2}$$

Theorem 2.1 (Lower complexity bound for strongly convex functions)

For any $x_0 \in \ell_2(\mathbb{N})$ and $\gamma, L > 0$, there exists a γ-strongly convex one times continuously differentiable function f with L-Lipschitz gradient such that for any algorithm satisfying (2.2), we have for all k

$$f(x_k) - f(x_*) \geqslant \frac{\gamma}{2}\left(\frac{\sqrt{Q}-1}{\sqrt{Q}+1}\right)^{2k} \|x_0 - x_*\|^2 .$$

where $Q = L/\gamma \geqslant 1$ is the condition number and x_* is a minimiser of f.

Accelerated gradient descent

Initialization $x_0 = \bar{x}_0 \in \mathrm{dom}(F)$ and $\lambda_0 = 0$

while *stopping criterion not satisfied* **do**

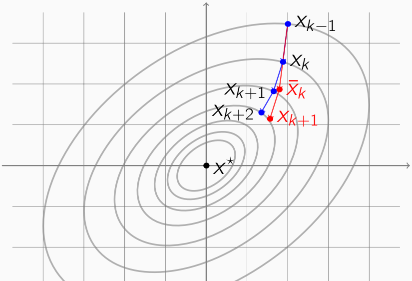
choose step-size $\gamma_k > 0$

Update $x_{k+1} = \bar{x}_k - \gamma_k \nabla F(\bar{x}_k)$

$\lambda_k = \frac{1+\sqrt{1+4\lambda_{k-1}^2}}{2}$ and $a_k = \frac{1-\lambda_k}{\lambda_{k+1}}$

$\bar{x}_{k+1} = x_{k+1} + a_k(x_{k+1} - x_k)$

end



If F is L-Lipschitz gradient, then by choosing $\gamma_k = 1/L$, AGD achieves $\mathcal{O}(1/k^2)$ convergence rate

Outline

Nonsmooth optimisation

Let $R : \mathbb{R}^n \to (-\infty, +\infty]$ be proper, convex and lower semi-continuous, but non-differentiable. Assume that $\operatorname{argmin}(R) \neq \emptyset$ and consider

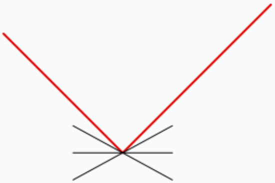
$$\min_{x \in \mathbb{R}^n} R(x). \tag{3.1}$$

Let $R : \mathbb{R}^n \to (-\infty, +\infty]$ be proper, convex and lower semi-continuous, but non-differentiable. Assume that $\operatorname{argmin}(R) \neq \emptyset$ and consider

$$\min_{x \in \mathbb{R}^n} R(x). \tag{3.1}$$

Recall: the subdifferential of R at x is the set

$$\partial R(x) = \{p \in \mathbb{R}^n \; ; \; R(y) - R(x) \geqslant \langle p, y - x \rangle, \quad \forall y \in \mathbb{R}^n\}.$$



$$\partial \|x\|_1 = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \\ [-1, 1] & x = 0 \end{cases}.$$

Note that $\partial R(x) = \{\nabla R(x)\}$ when R is differentiable at x.

Subgradient descent

Initialization $x_0 \in \mathrm{dom}(R)$

while *stopping criterion not satisfied* **do**

  choose step-size $\gamma_k > 0$ and a subgradient $g_k \in \partial R(x_k)$

  Update $x_{k+1} = x_k - \gamma_k g_k$

end

Subgradient descent

Initialization $x_0 \in \mathrm{dom}(R)$

while *stopping criterion not satisfied* **do**

  choose step-size $\gamma_k > 0$ and a subgradient $g_k \in \partial R(x_k)$

  Update $x_{k+1} = x_k - \gamma_k g_k$

end

In general, if x_* is a solution

$$\begin{aligned} \|x_{i+1} - x_*\|^2 &= \|x_i - x_*\|^2 + \|\gamma_i g_i\|^2 - 2\gamma_i \langle g_i, x_i - x_* \rangle \\ &\leqslant \|x_i - x_*\|^2 + \|\gamma_i g_i\|^2 - 2\gamma_i (R(x_i) - R(x_*)) \end{aligned}$$

So, summing from $i = 0, \ldots, k$ and rearranging, we have

$$\sum_{i=0}^{k} \gamma_i (R(x_i) - R(x_*)) \leqslant \|x_0 - x_*\|^2 + \sum_{i=0}^{k} \gamma_i^2 \|g_i\|^2 - \|x_{k+1} - x_*\|^2$$

Subgradient descent

Initialization $x_0 \in \mathrm{dom}(R)$

while *stopping criterion not satisfied* **do**

  choose step-size $\gamma_k > 0$ and a subgradient $g_k \in \partial R(x_k)$

  Update $x_{k+1} = x_k - \gamma_k g_k$

end

If R is L-Lipschitz,

$$2 \min_{i=0}^{k} (R(x_i) - R(x_*)) \leqslant \frac{\|x_0 - x_*\|^2 + L \sum_{i=0}^{k} \gamma_i^2}{2 \sum_{i=0}^{k} \gamma_i}$$

Subgradient descent

Initialization $x_0 \in \mathrm{dom}(R)$

while *stopping criterion not satisfied* **do**

choose step-size $\gamma_k > 0$ and a subgradient $g_k \in \partial R(x_k)$

Update $x_{k+1} = x_k - \gamma_k g_k$

end

If R is L-Lipschitz,

$$2 \min_{i=0}^{k} (R(x_i) - R(x_*)) \leqslant \frac{\|x_0 - x_*\|^2 + L \sum_{i=0}^{k} \gamma_i^2}{2 \sum_{i=0}^{k} \gamma_i}$$

- No convergence if γ_k is fixed. Note also that this is not a descent method.

Subgradient descent

Initialization $x_0 \in \mathrm{dom}(R)$

while *stopping criterion not satisfied* **do**

  choose step-size $\gamma_k > 0$ and a subgradient $g_k \in \partial R(x_k)$

  Update $x_{k+1} = x_k - \gamma_k g_k$

end

If R is L-Lipschitz,

$$2 \min_{i=0}^{k} (R(x_i) - R(x_*)) \leqslant \frac{\|x_0 - x_*\|^2 + L \sum_{i=0}^{k} \gamma_i^2}{2 \sum_{i=0}^{k} \gamma_i}$$

- No convergence if γ_k is fixed. Note also that this is not a descent method.
- In general choose $\sum_i \gamma_i^2 < \infty$ and $\sum_i \gamma_i = +\infty$, so γ_i converges to 0.

Subgradient descent

Initialization $x_0 \in \mathrm{dom}(R)$

while *stopping criterion not satisfied* **do**

choose step-size $\gamma_k > 0$ and a subgradient $g_k \in \partial R(x_k)$

Update $x_{k+1} = x_k - \gamma_k g_k$

end

If R is L-Lipschitz,

$$2 \min_{i=0}^{k} (R(x_i) - R(x_*)) \leqslant \frac{\|x_0 - x_*\|^2 + L \sum_{i=0}^{k} \gamma_i^2}{2 \sum_{i=0}^{k} \gamma_i}$$

- No convergence if γ_k is fixed. Note also that this is not a descent method.
- In general choose $\sum_i \gamma_i^2 < \infty$ and $\sum_i \gamma_i = +\infty$, so γ_i converges to 0.
- Choosing $\gamma_i \equiv C/\sqrt{k+1}$ for k iterations, then

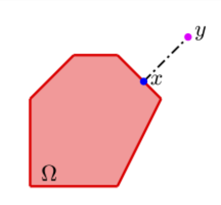
$$\min_{i=0}^{k} (R(x_i) - R(x_*)) \leqslant \frac{\|x_0 - x_*\|^2 + LC^2}{2C\sqrt{k+1}}.$$

Outline

Projection onto sets

Indicator function: Let $\Omega \subseteq \mathbb{R}^n$

$$\iota_\Omega(x) = \begin{cases} +\infty & x \notin \Omega \\ 0 & x \in \Omega. \end{cases}$$



Projection onto Ω:

$$\mathcal{P}_\Omega(y) \stackrel{\text{def.}}{=} \operatorname{argmin}_{x\in\Omega} \|x - y\| .$$

Constrained smooth optimisation

Let $F \in \mathcal{C}^1$ with L-Lipschitz gradient and let $\Omega \subset \mathbb{R}^n$ be a closed convex set, and consider

$$\min_{x \in \Omega} F(x). \tag{4.1}$$

Projected gradient descent

Constrained smooth optimisation

Let $F \in \mathcal{C}^1$ with L-Lipschitz gradient and let $\Omega \subset \mathbb{R}^n$ be a closed convex set, and consider

$$\min_{x \in \Omega} F(x). \tag{4.1}$$

Projected gradient descent method

Initialization $x_0 \in \Omega$

while *stopping criterion not satisfied* **do**

  choose step-size $\gamma_k \in (0, 2/L)$

  Update $x_{k+1/2} = x_k - \gamma_k \nabla F(x_k)$

  Project $x_{k+1} = \mathcal{P}_\Omega(x_{k+1/2})$

end

Examples

- Hyperplane: $\Omega = \{x : a^T x = b\},\ a \neq 0$

$$\mathcal{P}_\Omega = x + \frac{b - a^T x}{\|a\|^2} a.$$

- Affine subspace: $\Omega = \{x : Ax = b\}$ with $A \in \mathbb{R}^{m\times n},\ \operatorname{rank}(A) = m < n$

$$\mathcal{P}_\Omega = x + A^T (AA^T)^{-1}(b - Ax).$$

- Half space: $\Omega = \{x : a^T x \leqslant b\},\ a \neq 0$

$$\mathcal{P}_\Omega = x + \frac{b - a^T x}{\|a\|^2} a \quad \text{if } a^T x > b \qquad \text{and} \quad x \quad \text{if } a^T x \leqslant b.$$

- Nonnegative orthant: $\Omega = \mathbb{R}^n_+$

$$\mathcal{P}_\Omega = \left(\max\{0, x_i\}\right)_i.$$

Composite optimisation

$$\min_{x \in \mathbb{R}^n} \Phi(x) \stackrel{\text{def.}}{=} F(x) + R(x). \tag{4.2}$$

where $F \in \mathcal{C}^1$ has L-Lipschitz gradient and R is proper, convex, lower semi-continuous but nonsmooth.

Note that the constrained smooth problem can be rewritten as

$$\min_{x \in \mathbb{R}^n} F(x) + \iota_\Omega(x).$$

From projection to proximal mapping

Proximal mapping

Let R be proper, convex, lower semicontinuous and bounded from below. Its proximal mapping is defined by

$$\mathrm{prox}_{\gamma R}(y) \overset{\text{def.}}{=} \mathrm{argmin}_{x \in \mathbb{R}^n} \;\; \gamma R(x) + \frac{1}{2} \|x - y\|^2 .$$

Note that this is precisely the projection operator $\mathcal{P}_\Omega$ when $R = \iota_\Omega$.

Proximal mapping

Let R be proper, convex, lower semicontinuous and bounded from below. Its proximal mapping is defined by

$$\mathrm{prox}_{\gamma R}(y) \stackrel{\text{def.}}{=} \mathrm{argmin}_{x\in\mathbb{R}^n} \ \gamma R(x) + \frac{1}{2}\|x-y\|^2 .$$

Note that this is precisely the projection operator $\mathcal{P}_\Omega$ when $R = \iota_\Omega$.

Optimality condition denote $y_+ \stackrel{\text{def.}}{=} \mathrm{prox}_{\gamma R}(y)$,

$$\begin{aligned} 0 \in \gamma\partial R(y_+) + y_+ - y \quad &\Longleftrightarrow \quad y \in (\mathrm{Id} + \gamma\partial R)(y_+) \\ &\Longleftrightarrow \quad y_+ = (\mathrm{Id} + \gamma\partial R)^{-1}(y). \end{aligned}$$

Proximal gradient descent

Proximal gradient descent

Initialization $x_0 \in \mathrm{dom}(\Phi)$

while *stopping criterion not satisfied* **do**

  choose step-size $\gamma_k \in (0, 2/L)$

  Update $x_{k+1/2} = x_k - \gamma_k \nabla F(x_k)$

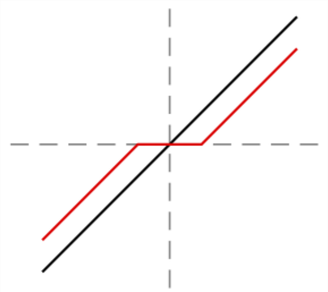
  Project $x_{k+1} = \mathrm{prox}_{\gamma_k R}(x_{k+1/2})$

end

Example

Soft-threshold: $R(x) = |x|$,

$$\operatorname{prox}_{\gamma R}(y) = \mathcal{T}_\gamma(y) = \begin{cases} y - \gamma : y > \gamma, \\ \quad 0 : y \in [-\gamma, \gamma], \\ y + \gamma : y < -\gamma. \end{cases}$$



Example

Consider the Lasso problem where for $y \in \mathbb{R}^m$ and a matrix $K \in \mathbb{R}^{n \times m}$,

$$R(x) = \lambda \|x\|_1 \quad \text{and} \quad F(x) = \frac{1}{2} \|Kx - y\|^2$$

- $\nabla F(x) = K^*(Kx - y)$
- $L = \|K^* K\|$
- The proximal mapping of γR is the soft-thresholding operator $\mathcal{T}_\gamma$

The iterates are therefore, for $\gamma_k \in (0, 1/\|K^* K\|]$:

$$x_{k+1} = \mathcal{T}_{\lambda \gamma_k}(x_k - \gamma_k(K^*(Kx_k - y)))$$

Specialised to the ℓ_1 case, this is sometimes known as the iterative soft thresholding algorithm (ISTA).

Interpretation

This is also known as Forward-Backward splitting:

$$x_{k+1} = \mathrm{prox}_{\gamma R}(x_k - \gamma \nabla F(x_k))$$

- forward: gradient descent set in F
- backward: implicit step in R

Interpretation

This is also known as Forward-Backward splitting:

$$x_{k+1} = \mathrm{prox}_{\gamma R}(x_k - \gamma \nabla F(x_k))$$

- forward: gradient descent set in F
- backward: implicit step in R

By definition of prox,

$$\begin{aligned} x_{k+1} &= \mathrm{argmin}_x \{ \frac{1}{2} \|x - (x_k - \gamma \nabla F(x_k))\|^2 + \gamma R(x) \} \\ &= \mathrm{argmin}_x \{ \frac{1}{2} \|x - x_k\|^2 + \gamma \langle x - x_k, \nabla F(x_k)) \rangle + \gamma R(x) \} \\ &= \mathrm{argmin}_x \{ F(x_k) + \langle x - x_k, \nabla F(x_k)) \rangle + \frac{1}{2\gamma} \|x - x_k\|^2 + R(x) \} \end{aligned}$$

So, x_{k+1} minimises $R(x)$ plus majorisation of $F(x)$ at x_k if $\gamma \leqslant \frac{1}{L}$

Convergence properties

FB is a descent method.

- For $\gamma_k \equiv \gamma \in (0, 1/L]$, x_k converges to a minimiser and

$$\Phi(x_k) - \Phi(x_*) \leqslant \frac{1}{2\gamma k} \|x_0 - x_*\|^2 .$$

- If F and R are strongly convex with parameters μ_F, μ_R and $\mu \stackrel{\text{def.}}{=} \mu_F + \mu_R > 0$, then

$$\Phi(x_k) - \Phi(x_*) \leqslant \omega^k \frac{(1 + \gamma\mu_R)}{2\gamma} \|x_0 - x_*\|^2 ,$$

  where $\omega = (1 - \gamma\mu_F)/(1 + \gamma\mu_R)$.

The convergence rate matches that of gradient descent, although faster convergence rates can be obtained by

- incorporating Nesterov acceleration. This is called FISTA (fast iterative soft thresholding).
- Another (very effective) acceleration technique for FB is the restarted-FISTA scheme.

Other examples of proximal operators

Quadratic function $R(x) = \frac{1}{2}x^T Ax + b^T x + c,\ A \succeq 0$

$$\mathrm{prox}_{\gamma R}(y) = (\mathrm{Id} + \gamma A)^{-1}(y - \gamma b).$$

Euclidean norm $R(x) = \|x\|$

$$\mathrm{prox}_{\gamma R}(y) = \begin{cases} (1 - \frac{\gamma}{\|y\|})y : \|y\| > \gamma, \\ 0 : o.w. \end{cases}$$

Nuclear norm $R(x) = \sum_i \sigma_i$

$$\mathrm{prox}_{\gamma R}(y) = U\mathcal{T}_{\gamma}\left(\Sigma\right)V^T.$$

Calculus rules for proximal operators

Quadratic perturbation $H(x) = R(x) + \frac{\alpha}{2} \|x\|^2 + \langle x,\, u \rangle + \beta,\ \alpha \geqslant 0$

$$\mathrm{prox}_H = \mathrm{prox}_{R/(\alpha+1)} \left(\frac{x - u}{\alpha + 1} \right).$$

Translation $H(x) = R(x - z)$

$$\mathrm{prox}_H = z + \mathrm{prox}_R(x - z).$$

Scaling $H(x) = R(x/\rho)$

$$\mathrm{prox}_H = \rho\, \mathrm{prox}_{R/\rho^2} \left(\frac{x}{\rho} \right).$$

Reflection $H(x) = R(-x)$

$$\mathrm{prox}_H = -\mathrm{prox}_R(-x).$$

Composition $H = R \circ L$ with L being bijective bounded linear mapping such that $L^{-1} = L^*$,

$$\mathrm{prox}_H = L^* \circ \mathrm{prox}_R \circ L.$$

Outline

We now let $F, G \in \Gamma_0(\mathbb{R}^n)$ and consider

$$\min_x F(x) + G(x) \tag{5.1}$$

Define the reflection operator:

$$\mathrm{rprox}_F \stackrel{\text{def.}}{=} 2\mathrm{prox}_F - \mathrm{Id}.$$

Douglas-Rachford splitting

Initialization $x_0 \in \mathrm{dom}(\Phi)$, $\gamma > 0$, $\mu \in (0, 2)$

while *stopping criterion not satisfied* **do**

$x_k = \mathrm{prox}_{\gamma F}(z_k)$

$z_{k+1} = (1 - \frac{1}{2}\mu) z_k + \frac{1}{2}\mu \, \mathrm{rprox}_{\gamma G}(\mathrm{rprox}_{\gamma F}(z_k))$

end

- There is guaranteed convergence to a fixed point: $z_k \to z_*$ if $\gamma > 0$ and $\mu \in (0, 2)$.
- This is a not a descent algorithm, and there are no known convergence rates for the general problem.

Basis Pursuit Let $F(x) = \|x\|_1$ and $G(x) = \iota_{\{x:Kx=b\}}$. Then, $\mathrm{prox}_{\gamma F}$ is the soft thresholding operator as before and $\mathrm{prox}_{\gamma G}(x) = x - A^\dagger(b - Ax)$.

Intuition for the Douglas-Rachford iterates

Note that the solution satisfies $0 \in \partial F(x_*) + \partial G(x_*)$.

So, there exists y_* such that

$$y_* \in \partial F(x_*) \quad \text{and} \quad -y_* \in \partial G(x_*).$$

Letting $z_* = y_* + z_*$, we can write for any $\gamma > 0$,

$$\begin{aligned}
& z_* - x_* \in \gamma \partial F(x_*) \quad \text{and} \quad x_* - z_* \in \gamma \partial G(x_*) \\
\iff & z_* \in x_* + \gamma \partial F(x_*) \quad \text{and} \quad 2x_* - z_* \in x_* + \gamma \partial G(x_*) \\
\iff & x_* = \operatorname{prox}_{\gamma F}(z_*) \quad \text{and} \quad 0 = \operatorname{prox}_{\gamma G}(2x_* - z_*) - x_* \\
\iff & x_* = \operatorname{prox}_{\gamma F}(z_*) \quad \text{and} \quad z_* = z_* + \mu \left(\operatorname{prox}_{\gamma G}(2x_* - z_*) - x_* \right).
\end{aligned}$$

Intuition for the Douglas-Rachford iterates

This leads to the fixed point iterations

$$\begin{aligned} x_k &= \mathrm{prox}_{\gamma F}(z_k) \\ z_{k+1} &= z_k + \mu \left(\mathrm{prox}_{\gamma G}(2x_k - z_k) - x_k \right) \end{aligned}$$

We can rewrite the z_k iterations as

$$\begin{aligned} z_{k+1} &= z_k + \mu \left(\mathrm{prox}_{\gamma G}(\mathrm{rprox}_{\gamma F}(z_k)) - \mathrm{prox}_{\gamma F}(z_k) \right) \\ &= (1 - \frac{\mu}{2}) z_k + \frac{\mu}{2} \, \mathrm{rprox}_{\gamma G}(\mathrm{rprox}_{\gamma F}(z_k)) \end{aligned}$$

ADMM

We now consider the following constrained optimisation problem:

$$\min_{x,y} F(x) + G(y) \quad \text{such that} \quad Ax + By = \zeta, \tag{5.2}$$

where $F, G \in \Gamma_0(\mathbb{R}^n)$.

The **augmented Lagrangian** formulation is

$$\min_{x,y} \sup_{\psi} F(x) + G(y) - \langle \psi,\, Ax + By - \zeta \rangle + \frac{\gamma}{2} \|Ax + By - z\|^2 \tag{5.3}$$

Note that the two formulations are equivalent since the supremum is $+\infty$ if $Ax + By \neq z$.

The ADMM iterations: alternate between descents on x, y and ascent on ψ

ADMM

Initialization $y_0, \psi_0, \gamma > 0$

while *stopping criterion not satisfied* **do**

$x_{k+1} \in \operatorname{argmin}_x F(x) + G(y_k) - \langle \psi_k, Ax + By_k - z \rangle + \frac{\gamma}{2} \|Ax + By_k - \zeta\|^2$

$y_{k+1} \in$
$\operatorname{argmin}_y F(x_{k+1}) + G(y) - \langle \psi, Ax_{k+1} + By - \zeta \rangle + \frac{\gamma}{2} \|Ax_{k+1} + By - \zeta\|^2$

$\psi_{k+1} = \psi_k + \gamma(\zeta - Ax_{k+1} - By_{k+1})$

end

Example

Consider the Lasso example again:

$$\min_{x,y} \lambda \|y\|_1 + \frac{1}{2}\|Kx - b\|^2 \text{ such that } x - y = 0.$$

The ADMM iterates are:

$$\begin{aligned} x_{k+1} &= (\mathrm{Id} + \gamma K^* K)^{-1}(K^* b - \gamma y_k - \psi_k) \\ y_{k+1} &= \mathrm{prox}_{\lambda/\gamma}(x_{k+1} + \psi_k/\gamma) \\ \psi_{k+1} &= \psi_k + \gamma(x_{k+1} - y_{k+1}). \end{aligned}$$

Equivalence to Douglas-Rachford

Suppose that F^* is continuous as A^*p and G^* is continuous at B^*q, then we can define

$$\tilde{F}(\xi) = \min_{Ax=\xi} F(x) \quad \text{and} \quad \tilde{G}(\eta) = \min_{By=\eta} G(y)$$

- the minimums are achieved (thanks to Fenchel Rockafellar duality).
- $\tilde{G}^*(q) = G^*(B^*q)$ and $\tilde{F}^*(p) = F^*(A^*p)$.

The Legendre-Fenchel dual of (4.1) is

$$\sup_p -F^*(A^*p) - G^*(B^*p) + \langle \zeta,\, p\rangle = -\min_p \tilde{F}^*(p) + \tilde{G}^*(p) - \langle \zeta,\, p\rangle.$$

We shall see that ADMM is equivalent to applying DR on this dual problem.

Equivalence to Douglas-Rachford

Rewrite ADMM iterates in terms of $\xi_k \stackrel{\text{def.}}{=} Ax_k$ and $\eta_k \stackrel{\text{def.}}{=} By_k$:

(i) $\xi_k = \operatorname{prox}_{\tilde{F}/\gamma}(\psi_k/\gamma + \zeta - \eta_k)$ since

$$(x_{k+1}, \xi_{k+1}) \in \operatorname{argmin}_{x,\xi : Ax=\xi} F(x) - \langle \psi_k,\, \xi \rangle + \frac{\gamma}{2} \|\xi + By_k - \zeta\|^2$$

$$\xi_{k+1} \in \operatorname{argmin}_{\xi} \tilde{F}(\xi) + \frac{\gamma}{2} \left\| \xi + By_k - \zeta - \frac{\psi_k}{\gamma} \right\|^2 \tag{5.4}$$

(ii) $\eta_{k+1} = \operatorname{prox}_{\tilde{G}/\gamma}(\psi_k/\gamma + \zeta - \xi_{k+1})$

Equivalence to Douglas-Rachford

By Moreau's identity $x = \mathrm{prox}_{f/\gamma}(x) + \frac{1}{\gamma}\mathrm{prox}_{\gamma f^*}(\gamma x)$:

(i) $\mathrm{prox}_{\gamma \tilde{F}^*}(\psi_k + \gamma\zeta - \gamma\eta_k) + \gamma\xi_{k+1} = \psi_k + \gamma(\zeta - \eta_k)$.

(ii) $\mathrm{prox}_{\gamma \tilde{G}^*}(\psi_k + \gamma(\zeta - \xi_{k+1})) = \psi_k + \gamma(\zeta - \xi_{k+1} - \eta_{k+1}) = \psi_{k+1}$.

Define

$$u_k \stackrel{\text{def.}}{=} \psi_k - \gamma\eta_k, \quad v_{k+1} \stackrel{\text{def.}}{=} \psi_k + \gamma(\zeta - \xi_{k+1}) \quad \text{and} \quad \tilde{F}^*_\zeta(p) \stackrel{\text{def.}}{=} \tilde{F}^*(p) - \langle \zeta,\, p\rangle.$$

Then

(i) $v_{k+1} = \gamma\eta_k + \mathrm{prox}_{\gamma \tilde{F}^*_\zeta}(u_k)$

(ii) From $\mathrm{prox}_{\gamma \tilde{G}^*}(v_{k+1}) = v_{k+1} - \gamma\eta_{k+1} = \psi_{k+1}$,

$$u_{k+1} = \mathrm{prox}_{\gamma \tilde{G}^*}(v_{k+1}) - \gamma\eta_{k+1} = 2\mathrm{prox}_{\gamma \tilde{G}^*}(v_{k+1}) - v_{k+1} = \mathrm{rprox}_{\gamma \tilde{G}^*}(v_{k+1})$$

The iterates are therefore

$$\begin{aligned} v_{k+1} &= \gamma\eta_k + \mathrm{prox}_{\gamma \tilde{F}^*_\zeta}(u_k) \\ u_{k+1} &= \mathrm{rprox}_{\gamma \tilde{G}^*}(v_{k+1}). \end{aligned}$$

This is precisely the Douglas-Rachford iterations

$$v_{k+1} = \frac{1}{2} v_k + \frac{1}{2}\mathrm{rprox}_{\gamma \tilde{F}^*_\zeta}(\mathrm{rprox}_{\gamma \tilde{G}^*}(v_k))$$

for the minimisation of $\min_p \Phi(p)$ where

$$\Phi(p) \stackrel{\text{def.}}{=} \tilde{F}^*_\zeta(p) + \tilde{G}^*(p) = \tilde{F}^*(p) + \tilde{G}^*(p) - \langle \zeta,\, p\rangle = F^*(A^*p) + G^*(B^*p) - \langle \zeta,\, p\rangle \tag{5.5}$$

which is precisely the dual formulation of (4.1).

Outline

Primal-Dual splitting

Consider the problem

$$\min_x F(Kx) + G(x). \tag{6.1}$$

Then, by considering the Fenchel conjugate of $F(Kx) = \sup_y \langle Kx, y\rangle - F^*(y)$, we have the saddle point problem

$$\min_x \sup_y \langle Kx, y\rangle - F^*(y) + G(x) \tag{6.2}$$

The primal dual splitting scheme is an implicit descent on x, followed by an implicit ascent on y: the optimality conditions are

$$x_* - \tau K^* y_* \in x_* + \tau \partial G(x_*) \quad \text{and} \quad y_* + \sigma K x_* \in y_* + \sigma \partial F^*(y_*)$$

Primal-Dual splitting

Primal-Dual splitting

Initialization $x_0, y_0,\ \sigma, \tau > 0$

while *stopping criterion not satisfied* **do**

$$x_{k+1} = \mathrm{prox}_{\tau G}(x_k - \tau K^* y_*)$$

$$y_{k+1} = \mathrm{prox}_{\sigma F^*}(y_k + \sigma K(2x_{k+1} - x_k))$$

end

Primal-Dual splitting

Initialization $x_0, y_0,\ \sigma, \tau > 0$

while *stopping criterion not satisfied* **do**

$$x_{k+1} = \mathrm{prox}_{\tau G}(x_k - \tau K^* y_*)$$

$$y_{k+1} = \mathrm{prox}_{\sigma F^*}(y_k + \sigma K(2x_{k+1} - x_k))$$

end

Example: Let D be the finite differences operator and $A \in \mathbb{R}^{m\times n}$.

$$\min_{x\in\mathbb{R}^n} \|Dx\|_1 + \frac{1}{2}\|Ax - b\|_2^2$$

We let $F(z) \stackrel{\text{def.}}{=} \|z\|_1$, $G(x) \stackrel{\text{def.}}{=} \frac{1}{2}\|Ax - b\|_2^2$ and $K \stackrel{\text{def.}}{=} D$.

We can compute $\mathrm{prox}_{\sigma F^*}$ easily using Moreau's identity and the soft thresholding operator. For $\mathrm{prox}_{\tau G}$, note that $z = \mathrm{prox}_{\tau G}(x)$ satisfies

$$z \in \mathrm{argmin}_z \frac{1}{2}\|z - x\|^2 + \frac{\tau}{2}\|Az - b\|^2 \iff z = (\mathrm{Id} + \tau A^* A)^{-1}(\gamma A^* b + x)$$

Primal-Dual splitting

Primal-Dual splitting

Initialization $x_0, y_0,\ \sigma, \tau > 0$

while *stopping criterion not satisfied* **do**

$$x_{k+1} = \mathrm{prox}_{\tau G}(x_k - \tau K^* y_*)$$

$$y_{k+1} = \mathrm{prox}_{\sigma F^*}(y_k + \sigma K(2x_{k+1} - x_k))$$

end

Convergence: If $\sigma\tau \left\|K\right\|^2 < 1$,

- (x_k, y_k) converges to a fixed point (x_*, y_*) which is a solution to the saddle point problem (6.2) if a solution exists.
- defining the primal-dual gap as

  $$\mathcal{G}(x,y) = \max_{y'}(\langle y',\ Kx\rangle - F^*(y') + G(x)) - \min_{x'}(\langle y,\ Kx'\rangle - F^*(y) + G(x')),$$

  we have $\mathcal{G}(\bar{x}_n, \bar{y}_n) \leqslant \frac{C}{k}$ where $\bar{x}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$ and $\bar{y}_n \overset{\text{def.}}{=} \frac{1}{n}\sum_{k=1}^{n} x_k$.

Moreau's identity

Let $F : \mathbb{R}^N \to (-\infty, \infty]$ be convex, proper, lower semi-continuous. Let F^* be its convex conjugate. Then, for all $\delta > 0$,

$$\mathrm{prox}_{\delta F}(x) + \delta \mathrm{prox}_{\delta^{-1} F^*}(x/\delta) = x.$$

Summary

There are two key ingredients to dealing with nonsmooth optimisation of the form

$$\min_x F(x) + \sum_i R_i(K_i x)$$

where F is smooth, R_i are non-smooth and K_i are linear operators.

1. Gradient descent
2. Proximal mapping

Key splitting methods:

- $F + R$ Forward-Backward splitting.
- $R_1 + R_2$ Douglas-Rachford splitting
- $R_1 + R_2(K\cdot)$ Primal-dual splitting, ADMM.