Convex Optimisation with the Alternating Directed Method of Multipliers Algorithm

Alex Gibberd

Lancaster University

29/10/2025

Overview

Background

Primal-Dual Optimality

Alternating Directed Method of Multipliers

Example: Piecewise Approximation

Summary

Background

Convex Optimisation[1]

We consider an optimisation problem

$$\min_{\theta} f(\theta)$$
$$\text{s.t. } g(\theta) \leq 0$$

to be convex if $f : \mathbb{R}^p \mapsto \mathbb{R}, g : \mathbb{R}^p \mapsto \mathbb{R}^q$ are convex functions:

- Recall function is convex if

$$f(\alpha x + \beta y) \leq \alpha f(x) + \beta f(y)$$

  for all $\alpha + \beta = 1$, $\alpha \geq 0, \beta \geq 0$ and $x, y \in \mathbb{R}^p$.
- We define the feasible set $\mathscr{F} := \{\theta \,|\, g(\theta) \leq 0\}$
- Define $\theta_\star$ to be a minimiser of the above problem

[1] See Convex Optimization by Boyd and Vandenberghe (esp. chapters 3, 5)

Global vs Local Optimisation

- **Global methods:**
  - Find the optimal $\theta_\star$ over all feasible points in $\mathscr{F}$
  - Worst case complexity grows exponentially with p, q
- **Local methods:**
  - Find an optimal $\tilde{\theta}_\star$ among feasible points that are deemed local.
  - We could define locality as points close to some neighbourhood of a starting point $\mathscr{F} \cup \mathscr{N}(\theta_0)$

Benefits of Convex Optimisation

Convexity gives us conditions for finding global opitma:

- Strict convexity implies there is a unique $\boldsymbol{\theta}_\star$
- Generally, we may have an equivalence class $\mathscr{F}_\star \subset \mathscr{F}$, such that

$$f(\boldsymbol{\theta}_\star \in \mathscr{F}_\star) < f(\boldsymbol{\theta} \in \mathscr{F} \setminus \mathscr{F}_\star)$$

  and within the equivalence class the objective is the same.

Benefits of Convex Optimisation

Convexity gives us conditions for finding global opitma:

- Strict convexity implies there is a unique $\boldsymbol{\theta}_\star$
- Generally, we may have an equivalence class $\mathscr{F}_\star \subset \mathscr{F}$, such that

$$f(\boldsymbol{\theta}_\star \in \mathscr{F}_\star) < f(\boldsymbol{\theta} \in \mathscr{F} \setminus \mathscr{F}_\star)$$

  and within the equivalence class the objective is the same.

Convex approximations (relaxations) can be useful as:

- Heuristics for Non-Convex problems ($\implies$ approximate solutions)
- A way to initialize complex non-convex problems, find a good starting point $\implies$ finding a good local minima?

Approach (for today)

- As formulated so far, we have both an objective to minimise, as well as constraints to consider.
- **Strategy:**
  - We will introduce some additional penalties, so that if we violate the constraints, a cost is added to the objective (i.e. the Lagrangian)
  - We can examine the saddle points of this augmented objective function
  - We will see that this gives rise to a related optimisation problem that is convex, even when the original problem is non-convex (weak-duality)

Approach (for today)

- As formulated so far, we have both an objective to minimise, as well as constraints to consider.
- **Strategy:**
  - We will introduce some additional penalties, so that if we violate the constraints, a cost is added to the objective (i.e. the Lagrangian)
  - We can examine the saddle points of this augmented objective function
  - We will see that this gives rise to a related optimisation problem that is convex, even when the original problem is non-convex (weak-duality)
  - ADMM is a structured way to solve this related problem (and also the original problem on the way).

Lagrangian

- We define the "primal" problem as

$$\begin{aligned} &\min_{\theta} f(\theta) \\ &\text{s.t. } g(\theta) \leq 0 \\ &\quad h(\theta) = 0\,, \end{aligned}$$

where $g(\cdot) \mapsto \mathbb{R}^q$, $h(\cdot) \mapsto \mathbb{R}^r$, and comparison is entrywise[2].

[2] For a convex problem we require $h(\theta)$ is affine, e.g. can be written $h(\theta) = \langle a, \theta \rangle - b$.

Lagrangian

- We define the "primal" problem as

$$\begin{aligned} &\min_{\theta} f(\theta) \\ &\text{s.t. } g(\theta) \leq 0 \\ &\quad h(\theta) = 0\,, \end{aligned}$$

  where $g(\cdot) \mapsto \mathbb{R}^q$, $h(\cdot) \mapsto \mathbb{R}^r$, and comparison is entrywise[2].

- We define the *Lagrangian* as:

$$L(\theta, v, v_e) := f(\theta) + \langle v, g(\theta) \rangle + \langle v_e, h(\theta) \rangle$$

  where $v \in \mathbb{R}^q, v_e \in \mathbb{R}^r$ are referred to as *Lagrange multipliers*.

[2] For a convex problem we require $h(\theta)$ is affine, e.g. can be written $h(\theta) = \langle a, \theta \rangle - b$.

Conjugate Function

The conjugate function $f^*()$ of $f()$ is defined as

$$f^*(y) := \sup_{x \in \mathrm{dom}(f)} \langle y, x \rangle - f(x)$$

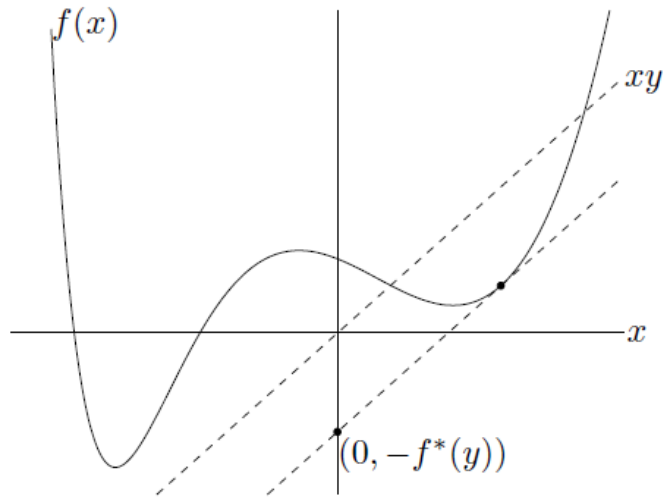

Figure 3.8 A function $f : \mathbf{R} \to \mathbf{R}$, and a value $y \in \mathbf{R}$. The conjugate function $f^*(y)$ is the maximum gap between the linear function yx and $f(x)$, as shown by the dashed line in the figure. If f is differentiable, this occurs at a point x where $f'(x) = y$.

Non-Convex $\rightarrow$ Convex

The conjugate function $f^\star$ is convex, even if f is not:

- This follows from definition, as a point-wise supremum over a family of affine (convex) functions.
- Taking this supremum preserves the convexity of the affine function.

Non-Convex $\rightarrow$ Convex

The conjugate function $f^\star$ is convex, even if f is not:

- This follows from definition, as a point-wise supremum over a family of affine (convex) functions.
- Taking this supremum preserves the convexity of the affine function.

When we work with the Lagrangian, we can be interested in seeing what the minimiser looks like as a function of the multipliers. This is defined as the dual function:

$$\begin{aligned} g(v, v_e) &= \inf_\theta L(\theta, v, v_e) \\ &= \inf_\theta \left(f(\theta) + \langle v, g(\theta) \rangle + \langle v_e, h(\theta) \rangle \right) \end{aligned}$$

- $g()$ is a point-wise infininum of affine functions, so is thus concave
- this is true even when $f()$ is not convex

Non-Convex $\rightarrow$ Convex

The dual, and conjugate functions are closely related. For example, consider

$$\min f(\theta) \quad s.t.\ \theta = 0\ .$$

We have[3]

$$\begin{aligned} L(\theta, v_e) &= f(\theta) + \langle v_e, \theta \rangle \\ g(v_e) &:= \inf_{\theta} \left[f(\theta) + \langle v_e, \theta \rangle \right] \\ &= -\sup_{\theta} \left[\langle -v_e, \theta \rangle - f(\theta) \right] \\ &= -f^*(-v_e)\ . \end{aligned}$$

[3] Recall the conjugate is defined as $f^*(\beta) := \sup_{\theta \in \mathrm{dom}(f)} \langle \beta, \theta \rangle - f(\theta)$.

Non-Convex $\rightarrow$ Convex

More generally, we may have

$$\begin{aligned} &\min f(\theta) \\ &\text{s.t. } A\theta \leq a \\ &\quad B\theta = b \end{aligned}$$

And thus[4]

$$\begin{aligned} L(\theta, v_e) &= f(\theta) + \langle v, A\theta - a \rangle + \langle v_e, B\theta - b \rangle \\ g(v_e) &:= \inf_\theta \left[f(\theta) + \langle v, A\theta - a \rangle + \langle v_e, B\theta - b \rangle \right] \\ &= -a^\top v - b^\top v_e + \inf_\theta \left[f(\theta) + (A^\top v + B^\top v_e)^\top \theta \right] \\ &= -a^\top v - b^\top v_e - f^*(-A^\top v - B^\top v_e) \end{aligned}$$

[4] Recall the conjugate is defined as $f^*(\beta) := \sup_{\theta \in \mathrm{dom}(f)} \langle \beta, \theta \rangle - f(\theta)$.

Primal-Dual Optimality

Lagrange Dual Problem

The dual function traces out the optimal value of the objective, as a function of the multipliers. Now, which value of multipliers is the best?

- This is known as the *Lagrangian Dual Problem*:

$$\begin{aligned} &\max g(v, v_e) \\ &\text{s.t. } v \geq 0 \end{aligned} \tag{1}$$

- We define a *dual-feasible pair*, as (v, v_e), such that for $v \geq 0$ we have $g(v, v_e) > -\infty$, i.e. v is feasible for the dual problem above.
- Define the *dual optimal pair* $(v^\star, v_e^\star)$ as the solution (maximiser) of (1)

Weak Duality

Let us define the optimal value of the dual problem and the primal problem

$$\mathscr{D} = g(v^\star, v_e^\star)$$
$$\mathscr{P} = f(\theta^\star)\,.$$

Then we have (even if the original problem is non-convex), that

$$\mathscr{D} \leq \mathscr{P}\,.$$

Weak Duality

Let us define the optimal value of the dual problem and the primal problem

$$\mathscr{D} = g(v^\star, v_e^\star)$$
$$\mathscr{P} = f(\theta^\star)\,.$$

Then we have (even if the original problem is non-convex), that

$$\mathscr{D} \leq \mathscr{P}\,.$$

Consequences:

- If primal problem is unbounded below such that $\mathscr{P} = -\infty$, then we have $\mathscr{D} = -\infty$
- If dual problem is unbounded above $\mathscr{D} = \infty$, then we also have $\mathscr{P} = \infty$ (which implies the primal problem is infeasible).

Verification

The fact $\mathscr{D} \leq \mathscr{P}$ can be easily verified:

- Let $\tilde{\theta} \in \mathscr{F}$ be a feasible point, i.e. $g(\tilde{\theta}) \leq 0$ and $h(\tilde{\theta}) = 0$
- This gives us the additional cost

$$\underbrace{\langle v, g(\tilde{\theta}) \rangle}_{\leq 0} + \underbrace{\langle v_e, h(\tilde{\theta}) \rangle}_{=0} \leq 0 \,,$$

  where we remember we have $v \geq 0$.
- Thus we have $L(\tilde{\theta}, v, v_e) \leq f(\tilde{\theta})$, for any $\tilde{\theta} \in \mathscr{F}$, and we recall $\theta_\star \in \mathscr{F}$.

Strong Duality

We refer to the difference $\text{gap} := \mathscr{P} - \mathscr{D} \geq 0$ as the *optimal duality gap*.

- If the primal problem is convex, then we usually have $\text{gap} = 0$
- If we have $\text{gap} = 0$ for a problem, then we say *strong-duality* holds.

[5] See Boyd Section 5.2. for examples and discussion.

Strong Duality

We refer to the difference $\text{gap} := \mathscr{P} - \mathscr{D} \geq 0$ as the *optimal duality gap*.

- If the primal problem is convex, then we usually have $\text{gap} = 0$
- If we have $\text{gap} = 0$ for a problem, then we say *strong-duality* holds.

Conditions under which we have strong-duality are known as *constraint qualifications*.

- If the conditions are met, then we can try to solve the dual problem, to solve the primal.
- Otherwise, solving the dual, can lower-bound the primal.
- Such conditions exist even for some (specific) non-convex problems[5].

[5] See Boyd Section 5.2. for examples and discussion.

Min-Max $\Longleftrightarrow$ Max-Min

- Consider the definition of the optimal values

$$\begin{aligned} \mathscr{P} &= \inf_{\theta} \sup_{v \geq 0} L(\theta, v) \\ \mathscr{D} &= \sup_{v \geq 0} \inf_{\theta} L(\theta, v) \end{aligned}$$

- Then

$$\mathscr{D} \leq \mathscr{P} \Longrightarrow \sup_{v \geq 0} \inf_{\theta} L(\theta, v) \leq \inf_{\theta} \sup_{v \geq 0} L(\theta, v)$$

- Strong-duality gives us

$$\sup_{v \geq 0} \inf_{\theta} L(\theta, v) = \inf_{\theta} \sup_{v \geq 0} L(\theta, v) \, .$$

Alternating Directed Method of Multipliers

Motivation[6]

Depending on the structure within the primal problem, solving the dual-problem may or may not be easier than the primal.

- We will focus on algorithms which iteratively try to solve the problem

$$\inf_{\theta} \sup_{v \geq 0} L(\theta, v) \implies \theta^\star = \arg\min_{\theta} L(\theta, v^\star) \tag{2}$$

- That is, we try a sequence of θ^k, v^k such that as $k \to \infty$ we have $\theta^k, v^k \to \theta^\star, v^\star$
- We assume strong-duality, so that we can use the implication (2).
- *We limit ourselves to convex problems*

[6] Recommend the review article by *Boyd et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers", 2011*

Dual Ascent

- Consider the equality constrained optimization

$$\min f(\theta)$$
$$\text{s.t. } A\theta = a\ .$$

- Assume that the dual $g(v)$ is differentiable, then letting

$$\theta' = \arg\min_{\theta} L(\theta, v)\ ,$$

gives us $\nabla g(v) = A\theta' - a$.

- This motivates a simple "algorithm" for finding the maximisers by iterating

$$\theta^{k+1} = \arg\min_{\theta} L(\theta, v^k)$$
$$v^{k+1} = v^k + \alpha(A\theta^{k+1} - a)$$

where $\alpha > 0$ is a step-size parameter.

Separability

- We will be particularly interested in the case where $f(\theta)$ is *linearly separable*, i.e.

$$f(\theta) = \sum_{i=1}^{m} f_i(\theta) \, .$$

- The equality constraint can also be partitioned $A = [A_1, \ldots, A_m]$ such that $A\theta = \sum_{i=1}^{m} A_i \theta_i$.

Separability

- We will be particularly interested in the case where $f(\theta)$ is *linearly separable*, i.e.

$$f(\theta) = \sum_{i=1}^{m} f_i(\theta) \, .$$

- The equality constraint can also be partitioned $A = [A_1, \ldots, A_m]$ such that $A\theta = \sum_{i=1}^{m} A_i \theta_i$.
- The Largrangian can then be written as

$$L(\theta, v) = \sum_{i=1}^{m} \left[f_i(\theta_i) + \langle v, A_i \theta_i \rangle - \frac{1}{m} \langle v, a \rangle \right] \, .$$

Separability

- We will be particularly interested in the case where $f(\theta)$ is *linearly separable*, i.e.

$$f(\theta) = \sum_{i=1}^{m} f_i(\theta) \, .$$

- The equality constraint can also be partitioned $A = [A_1, \ldots, A_m]$ such that $A\theta = \sum_{i=1}^{m} A_i \theta_i$.
- The Largrangian can then be written as

$$L(\theta, v) = \sum_{i=1}^{m} \left[f_i(\theta_i) + \langle v, A_i \theta_i \rangle - \frac{1}{m} \langle v, a \rangle \right] \, .$$

- This permits the splitting of the minimisation step, over $\{\theta_i\}_{i=1}^{m}$ as:

$$\theta_i^{k+1} = \arg\min_{\theta_i} L_i(\theta_i, v^k)$$

$$v^{k+1} = v^k + \alpha(A\theta^{k+1} - a)$$

Augmented Lagrangian

In order to enable convergence without strict convexity assumptions (or finiteness of f), researchers turned to the Augmented Lagrangian

$$L_\rho(\theta, v) = f(\theta) + \langle v, A\theta - a \rangle + \frac{\rho}{2} \|A\theta - a\|_2^2 .$$

- This adds curvature around the equality constraint $A\theta \approx a$
- $\rho > 0$ impacts the level of curvature added.
- Define the augmented dual as $g_\rho(v) := \inf_\theta L(\theta, v)$.

[7] The step-size ρ can be motivated through looking at the optimality (primal and dual feasibility) conditions and relating the update v^{k+1} with the previous step v^k, see e.g. p12 Boyd.

Augmented Lagrangian

In order to enable convergence without strict convexity assumptions (or finiteness of f), researchers turned to the Augmented Lagrangian

$$L_\rho(\theta, v) = f(\theta) + \langle v, A\theta - a\rangle + \frac{\rho}{2}\|A\theta - a\|_2^2 .$$

- This adds curvature around the equality constraint $A\theta \approx a$
- $\rho > 0$ impacts the level of curvature added.
- Define the augmented dual as $g_\rho(v) := \inf_\theta L(\theta, v)$.
- The *Method of Multipliers* algorithm is simply Lagrangian ascent applied to $g_\rho(v)$[7]

$$\theta^{k+1} = \arg\min_\theta L_\rho(\theta_i, v^k)$$
$$v^{k+1} = v^k + \rho(A\theta^{k+1} - a)$$

[7] The step-size ρ can be motivated through looking at the optimality (primal and dual feasibility) conditions and relating the update v^{k+1} with the previous step v^k, see e.g. p12 Boyd.

Alternating Directed Method of Multipliers

We now consider the setting where the objective is seperable, and the equality constraint links these components, i.e.

$$\min_{\theta,\beta} f_A(\theta) + f_B(\beta)$$
$$\text{s.t. } A\theta + B\beta = c$$

- The linear equality constraints allow us to let β be a linear transformation of θ.

Alternating Directed Method of Multipliers

We now consider the setting where the objective is seperable, and the equality constraint links these components, i.e.

$$\min_{\theta,\beta} f_A(\theta) + f_B(\beta)$$
$$\text{s.t. } A\theta + B\beta = c$$

- The linear equality constraints allow us to let β be a linear transformation of θ.
- E.g. In statistics, $-f_A(\theta)$ may be a (log)likelihood and $-f_B(T\theta)$ may be a log-prior, which leads to the minimiser being equivalent to the MAP.

Alternating Directed Method of Multipliers

We now consider the setting where the objective is seperable, and the equality constraint links these components, i.e.

$$\min_{\theta,\beta} f_A(\theta) + f_B(\beta)$$
$$\text{s.t. } A\theta + B\beta = c$$

- The linear equality constraints allow us to let β be a linear transformation of θ.
- E.g. In statistics, $-f_A(\theta)$ may be a (log)likelihood and $-f_B(T\theta)$ may be a log-prior, which leads to the minimiser being equivalent to the MAP.
- This also allows for the block-seperability as before, e.g. if $f(\theta) = \sum_i f_i(\theta_i)$

Alternating Directed Method of Multipliers

Let us write the augmented Lagrangian, and then consider the method of multipliers

$$L_\rho(\theta,\beta,v) = f_A(\theta) + f_B(\beta) + \langle v, \underbrace{A\theta + B\beta - c}_{=:\text{res}} \rangle + \frac{\rho}{2}\|A\theta + B\beta - c\|_2^2 .$$

Given the seperability of $f_A(\theta)$ and $f_B(\beta)$, we can consider $\arg\min_{(\theta,\beta)} L_\rho(\theta,\beta,v)$ in two distinct updates

$$\begin{aligned}
\theta^{k+1} &= \arg\min_\theta L_\rho(\theta,\beta^k,v^k) \\
\beta^{k+1} &= \arg\min_\beta L_\rho(\theta^{k+1},\beta,v^k) \\
v^{k+1} &= v^k + \rho(\underbrace{A\theta^{k+1} + B\beta^{k+1} - a}_{=:r^k})
\end{aligned}$$

Convergence (Theory)

Convergence of the iterates θ^k, β^k, v^k is guaranteed under relatively assumptions:

1. $f : \mathbb{R}^p \mapsto \mathbb{R} \cup \{+\infty\}$, $g : \mathbb{R}^q \mapsto \mathbb{R} \cup \{+\infty\}$ are closed, proper and convex.[8]
2. There exists $(\theta^\star, \beta^\star, v^\star)$ such that L_0 has a saddle point

$$L_0(\theta^\star, \beta^\star, v) \leq L(\theta^\star, \beta^\star, v^\star) \leq L(\theta, \beta, v^\star)$$

From which we obtain, $k \to \infty$:

- *Residual convergence*, $r^k \to 0$.
- *Objective convergence*, $f_A(\theta^k) + f_B(\beta^k) \to \mathscr{P}$
- *Dual convergence*, $v^k \to v^\star$, where $v^\star$ is a dual optimal point.

Further results are available under additional assumptions. [9]

[8] Implied iff the epigraphs, e.g. $\text{epi}(f)$ are closed and non-empty sets.

[9] See e.g. Wang 2019, *Global Convergence of ADMM in Nonconvex Nonsmooth Optimization*. Hong et al. 2017, *On the Linear Convergence of the Alternating Direction Method of Multipliers*.

Convergence (in practice)

The practical rate of convergence for ADMM varies on the application.

- This can be sensitive to the specification of $\rho \geq 0$.
- Some work considers adaptive step-size schemes where $\{\rho_k\}$ is a decreasing sequence.
- Usually, can converge to reasonable solutions in a few iterations (10's), but slow to converge to high-precision.

Convergence (in practice)

The practical rate of convergence for ADMM varies on the application.

- This can be sensitive to the specification of $\rho \geq 0$.
- Some work considers adaptive step-size schemes where $\{\rho_k\}$ is a decreasing sequence.
- Usually, can converge to reasonable solutions in a few iterations (10's), but slow to converge to high-precision.

It is worth to consider also the positives:

- Can easily be parallelised in the case that the θ^{k+1} or β^{k+1} updates are separable.
- Often, for popular choices of $f_A()$, $f_B()$ we may have closed form (efficient) solutions for the sub-problems (low iteration complexity).

Example: Piecewise Approximation

Trend-Filtering

- As a simple example, we will consider a variant of the "fused-Lasso".
- Basically, an ℓ_2 least-squares objective, with a penalty enforcing smoothness on the variation of the mean.
- Consider the model for time points $t = 1, \ldots, n$

$$y_t = \theta_t + \varepsilon_t \quad ; \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

$$\text{s.t.} \quad \sum_{t=2}^{n} |\theta_t - \theta_{t-1}| \leq \eta$$

and $\eta \geq 0$ represents a constraint on the variation of $\mathbb{E}[y_t] = \theta_t$.

Estimator (ℓ_1-Trend Filter)

We consider a penalised MLE

$$\arg\min_{\theta,\beta} \left[\underbrace{\frac{1}{2n}\|y-\theta\|_2^2}_{f_A(\theta)} + \underbrace{\lambda\|\beta\|_1}_{f_B(\beta)} \right]$$
$$\text{s.t. } \beta = D\theta \ ,$$

where $D \in \mathbb{R}^{n-1\times n}$ is a differencing matrix.

Estimator (ℓ_1-Trend Filter)

We consider a penalised MLE

$$\arg\min_{\theta,\beta} \left[\underbrace{\frac{1}{2n}\|y-\theta\|_2^2}_{f_A(\theta)} + \underbrace{\lambda\|\beta\|_1}_{f_B(\beta)} \right]$$
$$\text{s.t. } \beta = D\theta \ ,$$

where $D \in \mathbb{R}^{n-1\times n}$ is a differencing matrix.
Write the augmented Lagrangian problem

$$\arg \min_{A\theta+B\beta=c} \max_{v\in\mathbb{R}^n} \left[f_A(\theta) + f_B(\beta) + \langle v, \text{res}\rangle + \frac{\rho}{2}\|\text{res}\|_2^2 \right]$$

where we choose $A = D$, $B = -I$, $c = 0$

.

ADMM (ℓ_1-Trend Filter)

Recall the general problem

$$\arg \min_{A\theta+B\beta=c} \max_{v\in\mathbb{R}^n} \underbrace{\left[f_A(\theta) + f_B(\beta) + \langle v, \text{res} \rangle + \frac{\rho}{2} \|\text{res}\|_2^2 \right]}_{L_\rho(\theta,\beta,v)}$$

- With a simple rescaling, $\tilde{v} = v/\rho$ we can rewrite $L_\rho(\theta,\beta,v)$ as

$$\begin{aligned} L_\rho(\theta,\beta,v) &\equiv L_\rho(\theta,\beta,\tilde{v}) \\ &= f_A(\theta) + f_B(\beta) + \frac{\rho}{2}\|\text{res} + \tilde{v}\|_2^2 - \frac{\rho}{2}\|\tilde{v}\|_2^2 \end{aligned}$$

ADMM (ℓ_1-Trend Filter)

Recall the general problem

$$\arg \min_{A\theta+B\beta=c} \max_{v\in\mathbb{R}^n} \underbrace{\left[f_A(\theta) + f_B(\beta) + \langle v, \text{res} \rangle + \frac{\rho}{2} \|\text{res}\|_2^2 \right]}_{L_\rho(\theta,\beta,v)}$$

- With a simple rescaling, $\tilde{v} = v/\rho$ we can rewrite $L_\rho(\theta,\beta,v)$ as

$$\begin{aligned} L_\rho(\theta,\beta,v) &\equiv L_\rho(\theta,\beta,\tilde{v}) \\ &= f_A(\theta) + f_B(\beta) + \frac{\rho}{2}\|\text{res} + \tilde{v}\|_2^2 - \frac{\rho}{2}\|\tilde{v}\|_2^2 \end{aligned}$$

- Now choosing $A = D$, $B = -I$, $c = 0$, and replacing v with $\tilde{v}$ in our algorithm, we find

$$\begin{aligned} \theta^{k+1} &= \arg\min_\theta L_\rho(\theta, \beta^k, \tilde{v}^k) \\ \beta^{k+1} &= \arg\min_\beta L_\rho(\theta^{k+1}, \beta, \tilde{v}^k) \\ \tilde{v}^{k+1} &= \tilde{v}^k + D\theta^{k+1} - \beta^{k+1} \end{aligned}$$

ADMM (ℓ_1-Trend Filter)

Let us consider the specific updates involved:

1. *"Primal"* update for $\theta^{k+1} = \arg\min_\theta L_\rho(\theta, \beta^k, \tilde{v}^k)$

$$\begin{aligned}
\theta^{k+1} &= \arg\min_\theta \left[f_A(\theta) + \frac{\rho}{2} \|\text{res} + \tilde{v}^k\|_2^2 \right] \\
&= \arg\min_\theta \left[\frac{1}{2n} \|y - \theta\|_2^2 + \frac{\rho}{2} \|D\theta - \beta^{k+1} + \tilde{v}^k\|_2^2 \right] \\
&= (n^{-1} I_n + \rho D^\top D)^{-1} \left[\frac{y}{n} + \rho D^\top (\beta^k - \tilde{v}^k) \right]
\end{aligned}$$

ADMM (ℓ_1-Trend Filter)

Let us consider the specific updates involved:

1. *"Primal"* update for $\theta^{k+1} = \arg\min_\theta L_\rho(\theta, \beta^k, \tilde{v}^k)$

$$\begin{aligned}
\theta^{k+1} &= \arg\min_\theta \left[f_A(\theta) + \frac{\rho}{2} \|\text{res} + \tilde{v}^k\|_2^2 \right] \\
&= \arg\min_\theta \left[\frac{1}{2n} \|y - \theta\|_2^2 + \frac{\rho}{2} \|D\theta - \beta^{k+1} + \tilde{v}^k\|_2^2 \right] \\
&= (n^{-1} I_n + \rho D^\top D)^{-1} \left[\frac{y}{n} + \rho D^\top (\beta^k - \tilde{v}^k) \right]
\end{aligned}$$

1. *"Auxiliary"* update for $\beta^{k+1} = \arg\min_\beta L_\rho(\theta^{k+1}, \beta, \tilde{v}^k)$:

$$\begin{aligned}
\beta^{k+1} &= \arg\min_\theta \left[\lambda \|\beta\|_1 + \frac{\rho}{2} \|\text{res} + \tilde{v}^k\|_2^2 \right] \\
&= \arg\min_\theta \left[\frac{\rho}{2} \|D\theta^{k+1} + \tilde{v}^k - \beta\|_2^2 + \lambda \|\beta\|_1 \right] \\
&\equiv \text{prox}_{\frac{2\lambda}{\rho}\|\cdot\|_1}(D\theta^{k+1} + \tilde{v}^k) \\
&= \text{soft}(D\theta^{k+1} + \tilde{v}^k, 2\lambda/\rho)
\end{aligned}$$

Summary

Conclusion

- The ADMM algorithm is an evolution of basic Lagrangian ascent algorithms
- Idea, try to maximise the dual function to find a primal solution

Pros:

- Can easily make use of structure in common problems via known proximity operators
- Can be easily distributed/parallelised
- Guaranteed convergence under broad assumptions (convexity)
- Can be heuristically extended to allow for multiple blocks (e.g. multiple penalties)

Cons:

- Convergence rate is not necessarily great.
- Needs some thinking about how to split problem in most efficient way

Conclusion

- The ADMM algorithm is an evolution of basic Lagrangian ascent algorithms
- Idea, try to maximise the dual function to find a primal solution

Pros:

- Can easily make use of structure in common problems via known proximity operators
- Can be easily distributed/parallelised
- Guaranteed convergence under broad assumptions (convexity)
- Can be heuristically extended to allow for multiple blocks (e.g. multiple penalties)

Cons:

- Convergence rate is not necessarily great.
- Needs some thinking about how to split problem in most efficient way

Thank you for listening!