PART II: PRIMAL AND DUAL PROXIMAL GRADIENT TYPE METHODS

Preliminaries to Part II – Smoothness

Definition. Let $L \geq 0$. A function $f : \mathbb{E} \to (-\infty, \infty]$ is said to be L-smooth over a set $D \subseteq \operatorname{int}(\operatorname{dom}(f))$ if it is differentiable over D and satisfies

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \text{ for all } \mathbf{x}, \mathbf{y} \in D.$$

The constant L is called the smoothness parameter.

- The class of L-smooth functions is denoted by $C_L^{1,1}(D)$.
- When $D = \mathbb{E}$, the class is often denoted by $C_L^{1,1}$.
- The class of functions which are L-smooth for some $L \geq 0$ is denoted by $C^{1,1}$.
- If a function is L_1-smooth, then it is also L_2-smooth for any $L_2 \geq L_1$.

Examples:

- $f(\mathbf{x}) = \langle \mathbf{a}, \mathbf{x} \rangle + \mathbf{b}$, $\mathbf{a} \in \mathbb{E}$, $b \in \mathbb{R}$ (0-smooth).
- $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$, $\mathbf{A} \in \mathbb{S}^n, \mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$ ($\|\mathbf{A}\|_2$-smooth).
- $f(\mathbf{x}) = \frac{1}{2}d_C^2$ ($f : \mathbb{E} \to \mathbb{R}$) (1-smooth)

The Descent Lemma

Lemma. Let $f : \mathbb{E} \to (-\infty, \infty]$ be an L-smooth function ($L \geq 0$) over a given convex set D. Then for any $\mathbf{x}, \mathbf{y} \in D$,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2.$$

Proof.

- By the fundamental theorem of calculus: $f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt$.
- $f(\mathbf{y}) - f(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt$.
- Thus,

$$\begin{aligned} |f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| &= \left| \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt \right| \\ &\overset{(*)}{\leq} \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|_* \cdot \|\mathbf{y} - \mathbf{x}\| dt \\ &\leq \int_0^1 tL \|\mathbf{y} - \mathbf{x}\|^2 dt = \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2, \end{aligned}$$

L-Smoothness and Boundedness of the Hessian

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function over $\mathbb{R}^n$. Then for a given $L \geq 0$, the following two claims are equivalent:

(i) f is L-smooth.

(ii) $\|\nabla^2 f(\mathbf{x})\|_2 \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$.

Corollary. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable convex function over $\mathbb{R}^n$. Then f is L-smooth w.r.t. the l_2-norm iff $\lambda_{\max}(\nabla^2 f(\mathbf{x})) \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$.

Examples

- $f(\mathbf{x}) = \sqrt{1 + \|\mathbf{x}\|_2^2}$ $(f : \mathbb{R}^n \to \mathbb{R})$. 1-smooth w.r.t. to l_2.
- $f(\mathbf{x}) = \log\left(e^{x_1} + e^{x_2} + \ldots + e^{x_n}\right)$ $(f : \mathbb{R}^n \to \mathbb{R})$. 1-smooth w.r.t. l_2 and l_∞-norms.

Lecture 5 - The Proximal Gradient Method (PGM)

The Proximal Gradient Method aims to solve the composite model:

$$\text{(P)} \quad \min\{F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

(A) $g : \mathbb{E} \to (-\infty, \infty]$ is proper closed and convex.

(B) $f : \mathbb{E} \to (-\infty, \infty]$ is proper and closed; $\text{dom}(g) \subseteq \text{int}(\text{dom}(f))$ and f L_f-smooth over $\text{int}(\text{dom}(f))$.

(C) The optimal set of problem (P) is nonempty and denoted by X^*. The optimal value of the problem is denoted by F_{opt}.

Three prototype examples:

- **unconstrained smooth minimization** $(g \equiv 0)$
$$\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{E}\}$$

- **convex constrained smooth minimization** $(g = \delta_C, C \neq \emptyset$ closed convex$)$
$$\min\{f(\mathbf{x}) : \mathbf{x} \in C\}$$

- l_1 **regularized problems** $(\mathbb{E} = \mathbb{R}^n,\ g(x) \equiv \lambda\|x\|_1)$
$$\min\{f(\mathbf{x}) + \lambda\|\mathbf{x}\|_1 : \mathbf{x} \in \mathbb{R}^n\}$$

The Idea

Instead of minimizing directly

$$\min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x}) + g(\mathbf{x})$$

Approximate f by a regularized linear approximation of f while keeping g fixed.

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k) + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}^k\|^2 + g(\mathbf{x}) \right\}$$

$$\mathbf{x}^{k+1} = \operatorname*{argmin}_{\mathbf{x}} \left\{ g(\mathbf{x}) + \frac{1}{2t_k}\left\|\mathbf{x} - (\mathbf{x}^k - t_k\nabla f(\mathbf{x}^k))\right\|^2 \right\}$$

Proximal Gradient Method

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

Three Prototype Examples Contd.

- **Gradient Method** ($g = 0$, unconstrained minimization)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t_k \nabla f(\mathbf{x}^k)$$

- **Gradient Projection Method** ($g = \delta_C$, constrained convex minimization)

$$\mathbf{x}^{k+1} = P_C(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k))$$

- **Iterative Soft-Thresholding Algorithm (ISTA)** ($g = \|\cdot\|_1$):

$$\mathbf{x}^{k+1} = \mathcal{T}_{\lambda t_k}\left(\mathbf{x}^k - t_k \nabla f(\mathbf{x}^k)\right)$$

where $\mathcal{T}_\alpha(\mathbf{u}) = [|\mathbf{u}| - \alpha \mathbf{e}] \odot \mathrm{sgn}(\mathbf{u})$.

The Proximal Gradient Method

- We will take the stepsizes as $t_k = \frac{1}{L_k}$.

The Proximal Gradient Method

Initialization: pick $\mathbf{x}^0 \in \mathrm{int}(\mathrm{dom}(f))$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $L_k > 0$.

(b) set $\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k} g}\left(\mathbf{x}^k - \frac{1}{L_k}\nabla f(\mathbf{x}^k)\right)$.

- The general update step can be written as $\mathbf{x}^{k+1} = T_{L_k}^{f,g}(\mathbf{x}^k)$
- $T_L^{f,g} : \mathrm{int}(\mathrm{dom}(f)) \to \mathbb{E}$ is the prox-grad operator defined by

$$T_L^{f,g}(\mathbf{x}) \equiv \mathrm{prox}_{\frac{1}{L} g}\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right).$$

- When the identities of f and g will be clear from the context, we will often omit the superscripts f, g and write $T_L(\cdot)$ instead of $T_L^{f,g}(\cdot)$.

Stepsize Strategies in the Convex Case

When f is also convex, we will define two possible stepsize strategies for which

$$f(\mathbf{x}^{k+1}) \le f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L_k}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2.$$

- **constant.** $L_k = L_f$ for all k.
- **backtracking procedure B2.** The procedure requires two parameters (s, η), where $s > 0$ and $\eta > 1$. Define $L_{-1} = s$. At iteration k, L_k is set to be equal to L_{k-1}. Then, while

$$f(T_{L_k}(\mathbf{x}^k)) > f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), T_{L_k}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \frac{L_k}{2}\|T_{L_k}(\mathbf{x}^k) - \mathbf{x}^k\|^2,$$

we set $L_k := \eta L_k$. That is, L_k is chosen as $L_k = L_{k-1}\eta^{i_k}$, where i_k is the smallest nonnegative integer for which

$$f(T_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k)) \le f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), T_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k) - \mathbf{x}^k \rangle + \frac{L_k}{2}\|T_{L_{k-1}\eta^{i_k}}(\mathbf{x}^k) - \mathbf{x}^k\|^2.$$

Remarks

- **Monotonicity of PGM.**

$$F(\mathbf{x}^k) - F(\mathbf{x}^{k+1}) \geq \frac{L_k}{2}\|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2.$$

$O(1/k)$ Rate of Convergence of Proximal Gradient

Theorem. Suppose that f is convex. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with either a constant stepsize rule or the backtracking procedure B2. Then for any $\mathbf{x}^* \in X^*$ and $k \geq 0$,

$$F(\mathbf{x}^k) - F_{\rm opt} \leq \frac{\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k},$$

where $\alpha = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ if the backtracking rule is employed.

Iteration Complexity of Algorithms

- An ε-optimal solution of problem (P) is a vector $\bar{\mathbf{x}} \in \text{dom}(g)$ satisfying $F(\bar{\mathbf{x}}) - F_{\rm opt} \le \varepsilon$.
- In complexity analysis, the following question is asked: how many iterations are required to obtain an ε-optimal solution? meaning how many iterations are required to obtain the condition $F(\mathbf{x}^k) - F_{\rm opt} \le \varepsilon$
- Recall: $F(\mathbf{x}^k) - F_{\rm opt} \le \frac{\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{2k}$.

Theorem[$O(1/\varepsilon)$ complexity of PGM]. For any k satisfying

$$k \ge \left\lceil \frac{\alpha L_f R^2}{2\varepsilon} \right\rceil$$

it holds that $F(\mathbf{x}^k) - F_{\rm opt} \le \varepsilon$, where R is an upper bound on $\|\mathbf{x}^* - \mathbf{x}^0\|$ for some $\mathbf{x}^* \in X^*$.

The Proximal Point Method

Consider the problem

$$\min g(\mathbf{x})$$

g - proper closed and convex.
Employing PGM with $f \equiv 0$ leads to the proximal point method ($c > 0$ arbitrary)

Proximal Point Method

$$\mathbf{x}^{k+1} = \mathrm{prox}_{cg}(\mathbf{x}^k)$$

- **Result** $g(\mathbf{x}^k) - g_{\mathrm{opt}} \leq \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{2ck}$ for all k.

Is it an implementable method????

Preliminaries on Strong Convexity

Definition. A function $f : \mathbb{E} \to (-\infty, \infty]$ is called σ-strongly convex for a given $\sigma > 0$, if $\text{dom}(f)$ is convex and the following inequality holds for any $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$ and $\lambda \in [0, 1]$:

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) - \frac{1}{2}\sigma\lambda(1 - \lambda)\|\mathbf{x} - \mathbf{y}\|^2.$$

- A strongly convex function is also convex.
- If the underlying space is Euclidean, then f is σ-strongly convex iff $f(\cdot) - \frac{\sigma}{2}\|\cdot\|^2$ is convex.
- **Example:** $\mathbb{E} = \mathbb{R}^n$ is endowed with the l_2-norm, let $f : \mathbb{R}^n \to \mathbb{R}$ be given by

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c,$$

where $\mathbf{A} \in \mathbb{S}^n_{++}, \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}$.
f is $\lambda_{\min}(\mathbf{A})$-strongly convex.

Strong Convexity

Lemma. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a σ-stongly convex function ($\sigma > 0$), and let $g : \mathbb{E} \to (-\infty, \infty]$ be convex. Then $f + g$ is σ-strongly convex.

Proof.

- For any $\mathbf{x}, \mathbf{y} \in \text{dom}(f) \cap \text{dom}(g)$ and $\lambda \in [0, 1]$.

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) - \frac{\sigma}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2.$$

- $g(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda g(\mathbf{x}) + (1-\lambda)g(\mathbf{y})$.
- Adding the two inequalities, we obtain

$$(f+g)(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda(f+g)(\mathbf{x})+(1-\lambda)(f+g)(\mathbf{y})-\frac{\sigma}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2,$$

  showing that $f + g$ is σ-strongly convex.

Example: $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 + \delta_C(\mathbf{x})$ is 1-strongly convex, where $\mathbb{E}$ is Euclidean, $\emptyset \neq C \subseteq \mathbb{E}$ convex.

Existence and Uniqueness of Minimizers

Theorem. Let $f : \mathbb{E} \to (-\infty, \infty]$ be a proper closed and σ-strongly convex function ($\sigma > 0$). Then

(a) f has a unique minimizer.

(b) $f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^*\|^2$ for all $\mathbf{x} \in \text{dom}(f)$, where $\mathbf{x}^*$ is the unique minimizer of f.

Examples of Strongly Convex Functions

$f(\mathbf{x})$	$\mathrm{dom}(f)$	s.c. parameter	norm
$\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c$ $(\mathbf{A} \in \mathbb{S}^n_{++}, \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R})$	$\mathbb{R}^n$	$\lambda_{\min}(\mathbf{A})$	l_2
$\frac{1}{2}\|\mathbf{x}\|^2 + \delta_C(\mathbf{x})$ $(\emptyset \neq C \subseteq \mathbb{E}$ convex$)$	C	1	Euclidean
$-\sqrt{1 - \|\mathbf{x}\|_2^2}$	$B_{\|\cdot\|_2}[\mathbf{0}, 1]$	1	l_2
$\frac{1}{2}\|\mathbf{x}\|_p^2 \ \ (p \in (1,2])$	$\mathbb{R}^n$	$p-1$	l_p
$\sum_{i=1}^n x_i \log x_i$	Δ_n	1	l_2 or l_1

Linear Rate of Convergence of PGM – Strongly Convex Case

Theorem. Suppose that f is σ-strongly convex ($\sigma > 0$). Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the proximal gradient method with either a constant stepsize rule or backtracking procedure B2. Let

$$\alpha = \begin{cases} 1, & \text{constant stepsize,} \\ \max\left\{\eta, \frac{s}{L_f}\right\}, & \text{backtracking.} \end{cases}$$

Then for any $\mathbf{x}^* \in X$ and $k \geq 0$,

(a) $\|\mathbf{x}^{k+1} - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma}{\alpha L_f}\right) \|\mathbf{x}^k - \mathbf{x}^*\|^2$.

(b) $\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \left(1 - \frac{\sigma}{\alpha L_f}\right)^k \|\mathbf{x}^0 - \mathbf{x}^*\|^2$.

(c) $F(\mathbf{x}^{k+1}) - F_{\rm opt} \leq \frac{\alpha L_f}{2} \left(1 - \frac{\sigma}{\alpha L_f}\right)^{k+1} \|\mathbf{x}^0 - \mathbf{x}^*\|^2$.

Complexity of PGM - the Strongly Convex Case

A direct result of the rate analysis:

Theorem. For any $k \geq 1$ satisfying

$$k \geq \alpha\kappa \log\left(\frac{1}{\varepsilon}\right) + \alpha\kappa \log\left(\frac{\alpha L_f R^2}{2}\right),$$

it holds that $F(\mathbf{x}^k) - F_{\text{opt}} \leq \varepsilon$, where R is an upper bound on $\|\mathbf{x}^0 - \mathbf{x}^*\|$ and $\kappa = \frac{L_f}{\sigma}$.

Exercise 0

Consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \underbrace{\frac{\lambda_1}{2}\|\mathbf{x}\|_2^2 + \lambda_2\|\mathbf{x}\|_1}_{\text{elastic net}}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$ and $\lambda_1, \lambda_2 > 0$.

(a) Write the proximal gradient method with constant stepsize for solving the problem with the decomposition $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ and $g(\mathbf{x}) = \frac{\lambda_1}{2}\|\mathbf{x}\|_2^2 + \lambda_2\|\mathbf{x}\|_1$

(b) Same as (a) but with the decomposition $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \frac{\lambda_1}{2}\|\mathbf{x}\|_2^2, g(\mathbf{x}) = \lambda_2\|\mathbf{x}\|_1$.

(c) Write explicitly the efficiency estimates of each of the methods proposed in (a) and (b). Show that the two methods are actually the same.

Lecture 6 – FISTA (Fast Proximal Gradient Method)

- **The model:**

$$(P)\ \min_{\mathbf{x}\in\mathbb{E}} f(\mathbf{x}) + g(\mathbf{x})$$

- **Underlying Assumptions:**

(A) $g : \mathbb{E} \to (-\infty, \infty]$ is proper closed and convex.

(B) $f : \mathbb{E} \to \mathbb{R}$ is L_f-smooth and convex.

(C) The optimal set of (P) is nonempty and denoted by X^*. The optimal value of the problem is denoted by $F_{\rm opt}$.

- **The Idea:** instead of making a step of the form

$$\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k} g}\left(\mathbf{x}^k - \frac{1}{L_k}\nabla f(\mathbf{x}^k)\right)$$

we will consider a step of the form

$$\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k} g}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla f(\mathbf{y}^k)\right)$$

where $\mathbf{y}^k$ is a special linear combination of $\mathbf{x}^k, \mathbf{x}^{k-1}$

FISTA

FISTA

Input: $(f, g, \mathbf{x}^0)$, where f and g satisfy properties (A) and (B) and $\mathbf{x}^0 \in \mathbb{E}$.
Initialization: set $\mathbf{y}^0 = \mathbf{x}^0$ and $t_0 = 1$.
General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

```
(a) pick $L_k > 0$.
(b) set $\mathbf{x}^{k+1} = \mathrm{prox}_{\frac{1}{L_k} g}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla f(\mathbf{y}^k)\right)$.
(c) set $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.
(d) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.
```

- The dominant computational steps of the proximal gradient and FISTA methods are the same: one proximal computation and one gradient evaluation.

Stepsize Rules

- **constant.** $L_k = L_f$ for all k.
- **backtracking procedure B3.** The procedure requires two parameters (s, η), where $s > 0$ and $\eta > 1$. Define $L_{-1} = s$. At iteration k, L_k is set to be equal to L_{k-1}. Then, while

$$f(T_{L_k}(\mathbf{y}^k)) > f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k \rangle + \frac{L_k}{2}\|T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k\|^2,$$

  we set $L_k := \eta L_k$. In other words, the stepsize is chosen as $L_k = L_{k-1}\eta^{i_k}$, where i_k is the smallest nonnegative integer for which

$$\begin{aligned} f(T_{L_{k-1}\eta^{i_k}}(\mathbf{y}^k)) \quad \leq \quad & f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), T_{L_{k-1}\eta^{i_k}}(\mathbf{y}^k) - \mathbf{y}^k \rangle + \\ & \frac{L_k}{2}\|T_{L_{k-1}\eta^{i_k}}(\mathbf{y}^k) - \mathbf{y}^k\|^2. \end{aligned}$$

In both stepsize rules,

$$f(T_{L_k}(\mathbf{y}^k)) \leq f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k \rangle + \frac{L_k}{2}\|T_{L_k}(\mathbf{y}^k) - \mathbf{y}^k\|^2.$$

Remarks

- $\beta L_f \leq L_k \leq \alpha L_f$, where

$$\alpha = \begin{cases} 1, & \text{constant}, \\ \max\left\{\eta, \frac{s}{L_f}\right\}, & \text{backtracking}, \end{cases} \quad \beta = \begin{cases} 1, & \text{constant}, \\ \frac{s}{L_f}, & \text{backtracking}. \end{cases}$$

- Easy to show by induction that $t_k \geq \frac{k+2}{2}$ for all $k \geq 0$.

$O(1/k^2)$ rate of convergence of FISTA

Theorem. Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by FISTA with either a constant stepsize rule or the backtracking procedure B3. Then for any $\mathbf{x}^* \in X^*$ and $k \geq 1$,

$$F(\mathbf{x}^k) - F_{\rm opt} \leq \frac{2\alpha L_f \|\mathbf{x}^0 - \mathbf{x}^*\|^2}{(k+1)^2},$$

where $\alpha = 1$ in the constant stepsize setting and $\alpha = \max\left\{\eta, \frac{s}{L_f}\right\}$ if the backtracking rule is employed.

ISTA/FISTA

Consider the model

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) + \lambda\|\mathbf{x}\|_1,$$

- $\lambda > 0$
- $f : \mathbb{R}^n \to \mathbb{R}$ convex and L_f-smooth.

Iterative Shrinkage/Thresholding Algorithm (ISTA):

$$\mathbf{x}^{k+1} = \mathcal{T}_{\lambda/L_f}\left(\mathbf{x}^k - \frac{1}{L_f}\nabla f(\mathbf{x}^k)\right).$$

Fast Iterative Shrinkage/Thresholding Algorithm (ISTA):

(a) $\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_f}}\left(\mathbf{y}^k - \frac{1}{L_f}\nabla f(\mathbf{y}^k)\right)$.

(b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$.

(c) $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

l_1-Regularized Least Squares

Consider the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1,$$

- $\mathbf{A}\in\mathbb{R}^{m\times n}, \mathbf{b}\in\mathbb{R}^m$ and $\lambda > 0$.
- Fits (P) with $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$ and $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$.
- f is L_f-smooth with $L_f = \|\mathbf{A}^T\mathbf{A}\|_{2,2} = \lambda_{\max}(\mathbf{A}^T\mathbf{A})$.

ISTA: $\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_k}}\left(\mathbf{x}^k - \frac{1}{L_k}\mathbf{A}^T(\mathbf{A}\mathbf{x}^k - \mathbf{b})\right).$

FISTA:

(a) $\mathbf{x}^{k+1} = \mathcal{T}_{\frac{\lambda}{L_k}}\left(\mathbf{y}^k - \frac{1}{L_k}\mathbf{A}^T(\mathbf{A}\mathbf{y}^k - \mathbf{b})\right).$

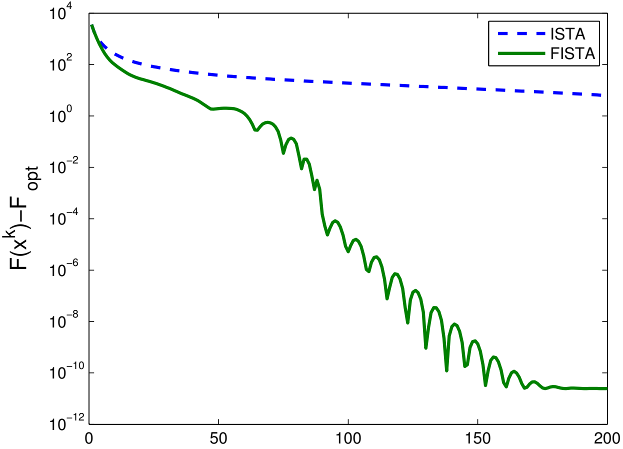
(b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}.$

(c) $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k).$

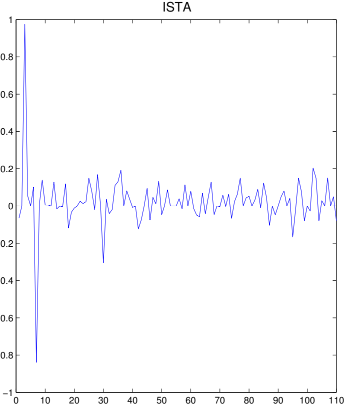
Numerical Example I

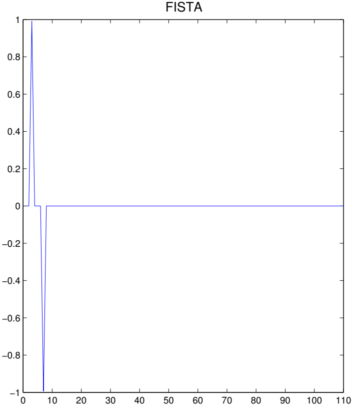
- test on regularized l_1-regularized least squares.
- $\lambda = 1$.
- $\mathbf{A} \in \mathbb{R}^{100\times 110}$. The components of $\mathbf{A}$ were independently generated using a standard normal distribution.
- the "true" vector is $\mathbf{x}_{\rm true} = \mathbf{e}_3 - \mathbf{e}_7$.
- $\mathbf{b} = \mathbf{A}\mathbf{x}_{\rm true}$.
- ran 200 iterations of ISTA and FISTA with $\mathbf{x}^0 = \mathbf{e}$.

Function Values



Solutions


ISTA
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0
10
20
30
40
50
60
70
80
90
100
110


FISTA
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
0
10
20
30
40
50
60
70
80
90
100
110

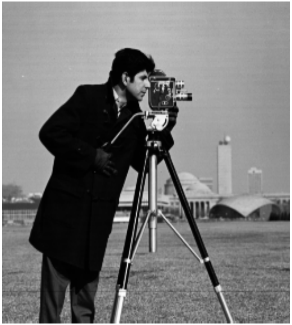
Example 2: Wavelet-Based Image Deblurring

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2+\lambda\|\mathbf{x}\|_1$$

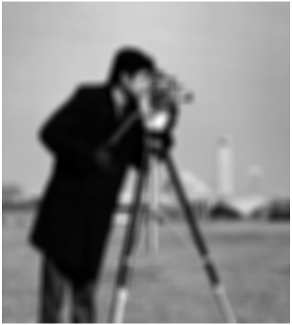
- image of size 512x512
- matrix **A** is dense (Gaussian blurring times inverse of two-stage Haar wavelet transform).
- all problems solved with fixed λ and Gaussian noise.

Deblurring of the Cameraman

original

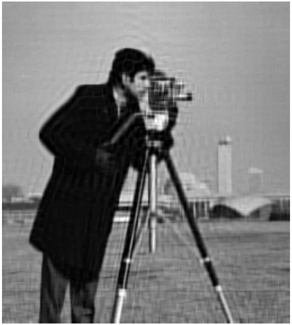


blurred and noisy

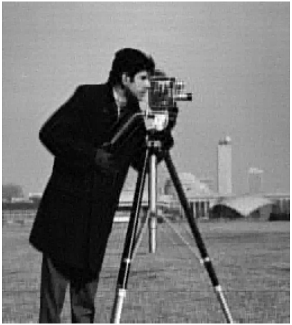


1000 Iterations of ISTA versus 200 of FISTA

ISTA: **1000 Iterations**

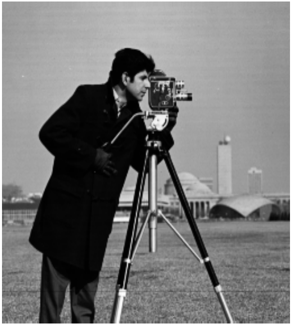


FISTA: **200 Iterations**

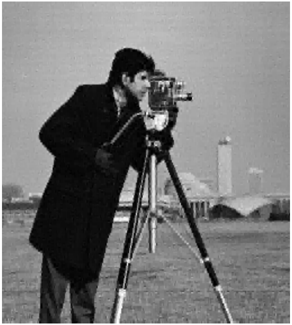


Original Versus Deblured via FISTA

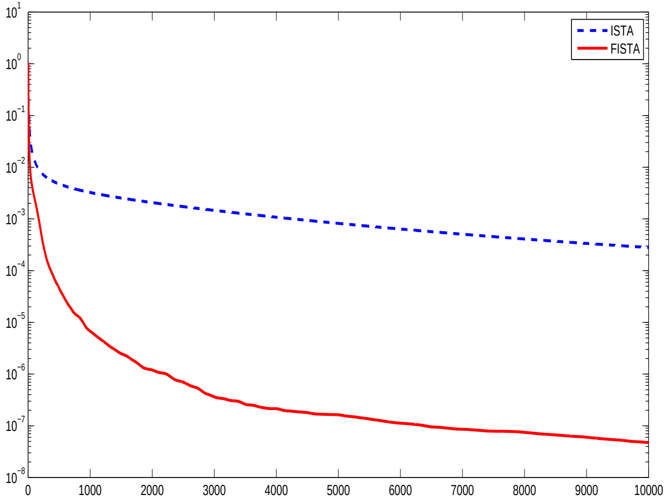
Original



FISTA:1000 Iterations



Function Values errors $F(\mathbf{x}^k) - F(\mathbf{x}^*)$



Restarting FISTA in the Strongly Convex Case

- Assume that f is σ-strongly convex for some $\sigma > 0$.
- The proximal gradient method attains an ε-optimal solution after an order of $O(\kappa \log(\frac{1}{\varepsilon}))$ iterations ($\kappa = \frac{L_f}{\sigma}$).
- A natural question is how the complexity result improves when using FISTA.
- Done by incorporating a restarting mechanism to FISTA – improves complexity result to $O(\sqrt{\kappa} \log(\frac{1}{\varepsilon}))$

Restarted FISTA

Initialization: pick $\mathbf{z}^{-1} \in \mathbb{E}$ and a positive integer N. Set $\mathbf{z}^0 = T_{L_f}(\mathbf{z}^{-1})$.

General step ($k \geq 0$)

- run N iterations of FISTA with constant stepsize ($L_k \equiv L_f$) and input $(f, g, \mathbf{z}^k)$ and obtain a sequence $\{\mathbf{x}^n\}_{n=0}^N$;
- set $\mathbf{z}^{k+1} = \mathbf{x}^N$.

Restarted FISTA

Theorem [$O(\sqrt{\kappa}\log(\frac{1}{\varepsilon}))$ complexity of restarted FISTA] Suppose that that f is σ-strongly convex ($\sigma > 0$). Let $\{\mathbf{z}^k\}_{k\geq 0}$ be the sequence generated by the restarted FISTA method employed with $N = \lceil\sqrt{8\kappa} - 1\rceil$. Let R be an upper bound on $\|\mathbf{z}^{-1} - \mathbf{x}^*\|$. Then

(a) $F(\mathbf{z}^k) - F_{\rm opt} \leq \frac{L_f R^2}{2}\left(\frac{1}{2}\right)^k$;

(b) after k iterations of FISTA with k satisfying

$$k \geq \sqrt{8\kappa}\left(\frac{\log(\frac{1}{\varepsilon})}{\log(2)} + \frac{\log(L_f R^2)}{\log(2)}\right),$$

an ε-optimal solution is obtained at the end of last completed cycle:

$$F(\mathbf{z}^{\lfloor\frac{k}{N}\rfloor}) - F_{\rm opt} \leq \varepsilon.$$

Sketch of Proof of (a)

(b) follows easily from (a). To prove (a),

- $F(\mathbf{z}^{n+1}) - F_{\rm opt} \leq \frac{2L_f\|\mathbf{z}^n - \mathbf{x}^*\|^2}{(N+1)^2}$.
- $F(\mathbf{z}^n) - F_{\rm opt} \geq \frac{\sigma}{2}\|\mathbf{z}^n - \mathbf{x}^*\|^2$.
- Hence, $F(\mathbf{z}^{n+1}) - F_{\rm opt} \leq \frac{4\kappa(F(\mathbf{z}^n) - F_{\rm opt})}{(N+1)^2}$.
- Since $N \geq \sqrt{8\kappa} - 1$, it follows that $\frac{4\kappa}{(N+1)^2} \leq \frac{1}{2}$, and hence,

$$F(\mathbf{z}^{n+1}) - F_{\rm opt} \leq \frac{1}{2}(F(\mathbf{z}^n) - F_{\rm opt}).$$

- $F(\mathbf{z}^k) - F_{\rm opt} \leq \left(\frac{1}{2}\right)^k (F(\mathbf{z}^0) - F_{\rm opt})$.
- Need to show that $F(\mathbf{z}^0) - F_{\rm opt} \leq \frac{L_f R^2}{2}$...

Second Attempt Handling Strong Convexity: V-FISTA

V-FISTA

Input: $(f, g, \mathbf{x}^0)$, where f and g satisfy properties (A) and (B) and $\mathbf{x}^0 \in \mathbb{E}$. f is σ-strongly convex.

Initialization: set $\mathbf{y}^0 = \mathbf{x}^0$ and $\kappa = \frac{L_f}{\sigma}$.

General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) pick $L_k > 0$.

(b) set $\mathbf{x}^{k+1} = \operatorname{prox}_{\frac{1}{L_k}g}\left(\mathbf{y}^k - \frac{1}{L_k}\nabla f(\mathbf{y}^k)\right)$.

(c) compute $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)(\mathbf{x}^{k+1} - \mathbf{x}^k)$.

Rate of convergence result:

$$F(\mathbf{x}^k) - F_{\mathrm{opt}} \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \left(F(\mathbf{x}^0) - F_{\mathrm{opt}} + \frac{\sigma}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|^2\right).$$

Exercise 0 Contd.

Suppose that $\mathbf{A}, \mathbf{b}, \lambda_1, \lambda_2$ are generated via

MATLAB

```
m=100;
n=120;
a = [0:m-1]+1;
b = [0:n-1]+0.5;
A = sin(10*(a'*b).^3);
xi = sin(31*[1:n].^3)';
b = A*xi;
lambda1=2;
lambda2=0.5;
```

Python

```
import numpy as np
m = 100
n = 120
a = np.arange(0,m)+1
b = np.arange(0,n)+0.5
A = np.sin(10 * np.outer(a,b)**3)
xi = np.sin(31 * np.arange(1,n+1)**3)
b = A @ xi
lambda1 = 2
lambda2 = 0.5
```

Implement the following methods in Python/MATLAB:

- proximal gradient with $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{\lambda_1}{2}\|\mathbf{x}\|_2^2, g(\mathbf{x}) = \lambda_2\|\mathbf{x}\|_1$.
- FISTA with $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{\lambda_1}{2}\|\mathbf{x}\|_2^2, g(\mathbf{x}) = \lambda_2\|\mathbf{x}\|_1$.
- V-FISTA with $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \frac{\lambda_1}{2}\|\mathbf{x}\|_2^2, g(\mathbf{x}) = \lambda_2\|\mathbf{x}\|_1$.
- V-FISTA with $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2, g(\mathbf{x}) = \frac{\lambda_1}{2}\|\mathbf{x}\|_2^2 + \lambda_2\|\mathbf{x}\|_1$. (yes - not a "legal" method. Use λ_1 as the strong-convexity parameter)

Run 100 iterations of each of the methods. Start each of the methods with the all-zeros vectors and use a constant stepsize. Plot the values of $F(\mathbf{x}^k) - F_{\rm opt}$ of the four methods (in the same plot). Use log-scale in the y-axis. Which of the methods performed best? write the first four components of the solution generated by the methods.

Exercise 1

Consider the problem

$$\min \sqrt{\mathbf{x}^T\mathbf{Q}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c} + 0.2\|\mathbf{D}\mathbf{x} + \mathbf{e}\|_1,$$

where $\mathbf{Q} \in \mathbb{R}^{30\times 30}, \mathbf{b} \in \mathbb{R}^{30}, c \in \mathbb{R}, \mathbf{D} \in \mathbb{R}^{10\times 30}$. The matrix $\mathbf{Q}$ is positive definite.

(a) Show that under the condition $c > \mathbf{b}^T\mathbf{Q}^{-1}\mathbf{b}$, the problem is well-defined (the expression inside the square root is always nonnegative)

(b) Show that under the above condition, the problem is convex.

(c) Assume that $\mathbf{D}\mathbf{D}^T = \mathbf{I}$. Write explicitly the proximal gradient method and FISTA with constant stepsize $1/L_f$ for solving the problem (taking $\|\mathbf{D}\mathbf{x}\|_1$ as the nonsmooth part). Explain how to compute a Lipschitz constant of the differentiable part.

Exercise 1 Contd.

(d) Generate $\mathbf{Q}, \mathbf{b}, c$ and $\mathbf{D}$ by the following commands:

MATLAB

```
n=30;
www=reshape([1:n^2],n,n);
A = sin(93*www.^3);
Q=A’*A;
b=10*sin(27*[1:n].^3)’;
c=b’*(Q\b)+1;
DD = sin(15*www.^3);
Z = sqrtm(inv(DD*DD’));
D = Z*DD;
```

Python

```
import numpy as np
import numpy.linalg as la
from scipy.linalg import sqrtm
n = 30
w = np.arange(1,n**2+1)
www = w.reshape((n,n)).T
A = np.sin(93*www**3)
Q = A.T @ A
b = 10*np.sin(27*np.arange(1,n+1)**3)
c = b @ la.solve(Q,b)+1
DD = np.sin(15*www**3)
Z = sqrtm(la.inv(DD@DD.T))
D = Z @ DD
```

Implement the methods from part (c). Run 1001 iterations of each of the methods. Start each of the methods with the all-ones vector. Print information on iterations 1,101,201,...,1001 (only one line for each iteration!). What is the stepsize that was used for the proximal gradient method? What is the vector found by each of the methods? (write explicitly the 30 values). Add a plot of $f(\mathbf{x}^k) - f_{\rm opt}$ of the two methods (in the same plot).

Lecture 7 – Dual-Based Proximal Gradient Methods

Main Model:

$$(P) \quad f_{\rm opt} = \min_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + g(\mathcal{A}(\mathbf{x})) \right\},$$

Underlying Assumptions:

(A) $f : \mathbb{E} \to (-\infty, +\infty]$ is proper closed and σ-strongly convex ($\sigma > 0$).

(B) $g : \mathbb{V} \to (-\infty, +\infty]$ is proper closed and convex.

(C) $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ is a linear transformation.

(D) there exists $\hat{\mathbf{x}} \in \text{ri}(\text{dom}(f))$ and $\hat{\mathbf{z}} \in \text{ri}(\text{dom}(g))$ such that $\mathcal{A}(\hat{\mathbf{x}}) = \hat{\mathbf{z}}$.

Existence and uniqueness of optimal solution: under the above assumptions, the objective function is proper closed and strongly convex, and hence there exists a unique optimal solution, which will be denoted by $\mathbf{x}^*$.

Example 1: Orthogonal Projection onto a Polyhedral set

- Let
$$S = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \le \mathbf{b}\},$$
where $\mathbf{A} \in \mathbb{R}^{p \times n}$, $\mathbf{b} \in \mathbb{R}^p$. Assume that $S \neq \emptyset$.
- Let $\mathbf{d} \in \mathbb{R}^n$. The orthogonal projection of $\mathbf{d}$ onto S is the unique optimal solution of
$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2 : \mathbf{A}\mathbf{x} \le \mathbf{b} \right\}.$$
- Fits model (P) with $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^p$, $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{d}\|^2$,
$$g(\mathbf{z}) = \delta_{\mathrm{Box}[-\infty\mathbf{e},\mathbf{b}]}(\mathbf{z}) = \begin{cases} \mathbf{0}, & \mathbf{z} \le \mathbf{b}, \\ \infty, & \text{else.} \end{cases}$$
and $\mathcal{A}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x}$.
- $\sigma = 1$

Example 2: One-Dimensional Total Variation Denoising

- **Denoising problem:**

$$\min_{\mathbf{x}\in\mathbb{E}} \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2 + R(\mathcal{A}(\mathbf{x})).$$

  - $\mathbf{d}\in\mathbb{E}$ - noisy and known signal
  - $\mathcal{A}:\mathbb{E}\to\mathbb{V}$ - linear transformation.
  - $R:\mathbb{V}\to\mathbb{R}_+$ - regularizing function measuring the magnitude of its argument.
- One-dimensional total variation denoising problem, $\mathbb{E}=\mathbb{R}^n, \mathbb{V}=\mathbb{R}^{n-1}, \mathcal{A}(\mathbf{x})=\mathbf{D}\mathbf{x}, R(\mathbf{z})=\lambda\|\mathbf{z}\|_1(\lambda>0)$, $\mathbf{D}$ defined by $\mathbf{D}\mathbf{x}=(x_1-x_2,x_2-x_3,\ldots,x_{n-1}-x_n)^T$

$$(P_1)\quad \min_{\mathbf{x}\in\mathbb{R}^n}\left\{\frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2+\lambda\|\mathbf{D}\mathbf{x}\|_1\right\}.$$

- More explicitly: $\min_{\mathbf{x}\in\mathbb{E}}\left\{\frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2+\lambda\sum_{i=1}^{n-1}|x_i-x_{i+1}|\right\}$.
- The function $\mathbf{x}\mapsto\|\mathbf{D}\mathbf{x}\|_1$ is a one-dimensional total variation function.
- Fits model (P) with $\mathbb{E}=\mathbb{R}^n, \mathbb{V}=\mathbb{R}^{n-1}, f(\mathbf{x})=\frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2(\sigma=1), g(\mathbf{y})=\lambda\|\mathbf{y}\|_1$, $\mathcal{A}(\mathbf{x})\equiv\mathbf{D}\mathbf{x}$

The Dual Problem

- (P) is the same as $\min_{\mathbf{x},\mathbf{z}}\{f(\mathbf{x})+g(\mathbf{z}):\mathcal{A}(\mathbf{x})-\mathbf{z}=\mathbf{0}\}$
- Lagrangian:
  $L(\mathbf{x},\mathbf{z};\mathbf{y})=f(\mathbf{x})+g(\mathbf{z})-\langle \mathbf{y},\mathcal{A}(\mathbf{x})-\mathbf{z}\rangle=f(\mathbf{x})+g(\mathbf{z})-\langle \mathcal{A}^T(\mathbf{y}),\mathbf{x}\rangle+\langle \mathbf{y},\mathbf{z}\rangle.$
- Minimizing the Lagrangian w.r.t. **x** and **z**, we obtain the dual problem

$$\boxed{(D)\quad q_{\rm opt}=\max_{\mathbf{y}\in\mathbb{V}}\left\{q(\mathbf{y})\equiv -f^*(\mathcal{A}^T(\mathbf{y}))-g^*(-\mathbf{y})\right\}.}$$

Theorem [strong duality of the pair (P),(D)] $f_{\rm opt}=q_{\rm opt}$ and the dual problem (D) attains an optimal solution.

The dual problem in minimization form:

$$\boxed{(D')\quad \min_{\mathbf{y}\in\mathbb{V}}\{F(\mathbf{y})+G(\mathbf{y})\}}$$

$$\begin{aligned}F(\mathbf{y}) &\equiv f^*(\mathcal{A}^T(\mathbf{y})),\\ G(\mathbf{y}) &\equiv g^*(-\mathbf{y}).\end{aligned}$$

The Conjugate Correspondence Theorem

Theorem [Conjugate Correspondence Theorem] Let $\sigma > 0$. Then

(a) If $f : \mathbb{E} \to \mathbb{R}$ is a $\frac{1}{\sigma}$-smooth convex function, then f^* is σ-strongly convex.

(b) If $f : \mathbb{E} \to (-\infty, \infty]$ is a proper closed σ-strongly convex function, then $f^* : \mathbb{E} \to \mathbb{R}$ is $\frac{1}{\sigma}$-smooth.

The Dual Problem

$$(D') \quad \min_{\mathbf{y}\in\mathbb{V}}\{F(\mathbf{y}) + G(\mathbf{y})\}$$

Properties of F and G:

(a) $F : \mathbb{V} \to \mathbb{R}$ is convex and L_F-smooth with $L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$;

(b) $G : \mathbb{V} \to (-\infty, \infty]$ is proper closed and convex.

Only non-trivial property is the L_F-smoothness of F: by the conjugate correspondence theorem f^* is $\frac{1}{\sigma}$-smooth. Therefore,

$$\begin{aligned}
\|\nabla F(\mathbf{y}_1) - \nabla F(\mathbf{y}_2)\| &= \|\mathcal{A}(\nabla f^*(\mathcal{A}^T(\mathbf{y}_1))) - \mathcal{A}(\nabla f^*(\mathcal{A}^T(\mathbf{y}_2)))\| \\
&\leq \|\mathcal{A}\| \cdot \|\nabla f^*(\mathcal{A}^T(\mathbf{y}_1)) - \nabla f^*(\mathcal{A}^T(\mathbf{y}_2))\| \\
&\leq \frac{1}{\sigma}\|\mathcal{A}\| \cdot \|\mathcal{A}^T(\mathbf{y}_1) - \mathcal{A}^T(\mathbf{y}_2)\| \\
&\leq \frac{\|\mathcal{A}\| \cdot \|\mathcal{A}^T\|}{\sigma}\|\mathbf{y}_1 - \mathbf{y}_2\| = \frac{\|\mathcal{A}\|^2}{\sigma}\|\mathbf{y}_1 - \mathbf{y}_2\|,
\end{aligned}$$

Dual Proximal Gradient

Dual Proximal Gradient = Proximal Gradient on (D')

Dual Proximal Gradient – dual representation

- **Initialization:** pick $\mathbf{y}^0 \in \mathbb{V}$ and $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$.
- **General step** $(k \geq 0)$**:**

$$\mathbf{y}^{k+1} = \operatorname{prox}_{\frac{1}{L}G}\left(\mathbf{y}^k - \frac{1}{L}\nabla F(\mathbf{y}^k)\right).$$

Theorem [rate of convergence of the dual objective function] Let $\{\mathbf{y}^k\}_{k\geq 0}$ be the sequence generated by the DPG method with $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. Then for any dual optimal solution $\mathbf{y}^*$ $k \geq 1$,

$$q_{\text{opt}} - q(\mathbf{y}^k) \leq \frac{L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{2k}.$$

Constructing a Primal Representation–Technical Lemma

Lemma. Let $F(\mathbf{y}) = f^*(\mathcal{A}^T(\mathbf{y}) + \mathbf{b}), G(\mathbf{y}) = g^*(-\mathbf{y})$, where f, g and $\mathcal{A}$ satisfy properties (A),(B) and (C) and $\mathbf{b} \in \mathbb{E}$. Then for any $\mathbf{y}, \mathbf{v} \in \mathbb{V}$ and $L > 0$ the relation

$$\mathbf{y} = \operatorname{prox}_{\frac{1}{L}G}\left(\mathbf{v} - \frac{1}{L}\nabla F(\mathbf{v})\right) \tag{1}$$

holds if and only if

$$\mathbf{y} = \mathbf{v} - \frac{1}{L}\mathcal{A}(\tilde{\mathbf{x}}) + \frac{1}{L}\operatorname{prox}_{Lg}(\mathcal{A}(\tilde{\mathbf{x}}) - L\mathbf{v}),$$

where

$$\tilde{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}}\left\{\langle \mathbf{x}, \mathcal{A}^T(\mathbf{v}) + \mathbf{b}\rangle - f(\mathbf{x})\right\}.$$

Dual Proximal Gradient - Primal Representation

The Dual Proximal Gradient (DPG) Method – primal representation

Initialization: pick $\mathbf{y}^0 \in \mathbb{V}$, and $L \geq \frac{\|\mathcal{A}\|^2}{\sigma}$.

General step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

(a) set $\mathbf{x}^k = \operatorname*{argmax}_{\mathbf{x}} \left\{ \langle \mathbf{x}, \mathcal{A}^T(\mathbf{y}^k) \rangle - f(\mathbf{x}) \right\}$;

(b) set $\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{L}\mathcal{A}(\mathbf{x}^k) + \frac{1}{L}\mathrm{prox}_{Lg}(\mathcal{A}(\mathbf{x}^k) - L\mathbf{y}^k)$.

- The sequence $\{\mathbf{x}^k\}_{k\geq 0}$ generated by the method will be called "the primal sequence", although its elements are not necessarily feasible.

The Primal-Dual Relation

Obtaining a rate of the primal sequence is done using the following result.

Lemma [primal-dual relation] Let $\bar{\mathbf{y}} \in \text{dom}(G)$, and let

$$\bar{\mathbf{x}} = \underset{\mathbf{x}\in\mathbb{E}}{\text{argmax}} \left\{ \langle \mathbf{x}, \mathcal{A}^T(\bar{\mathbf{y}}) \rangle - f(\mathbf{x}) \right\}.$$

Then

$$\|\bar{\mathbf{x}} - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\text{opt}} - q(\bar{\mathbf{y}})).$$

$O(1/k)$ Rate of the Primal Sequence Generated by DPG

Theorem. Let $\{\mathbf{x}^k\}_{k\geq 0}$ and $\{\mathbf{y}^k\}_{k\geq 0}$ be the primal and dual sequences generated by the DPG method with $L \geq L_F$. Then for any optimal dual solution $\mathbf{y}^*$ and $k \geq 1$,

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{\sigma k}.$$

Proof.

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\rm opt} - q(\mathbf{y}^k)) \leq \frac{2}{\sigma}\frac{L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{2k},$$

Fast Dual Proximal Gradient (FDPG)

Fast Dual Proximal Gradient = FISTA on (D')

Fast Dual Proximal Gradient (FDPG) - dual representation

- **Initialization:** $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}, t_0 = 1.$
- **General Step ($k \geq 0$):**
  (a) $\mathbf{y}^{k+1} = \mathrm{prox}_{\frac{1}{L}G}\left(\mathbf{w}^k - \frac{1}{L}\nabla F(\mathbf{w}^k)\right);$
  (b) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2};$
  (c) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k-1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k).$

Theorem [rate of convergence of the dual objective function] Let $\{\mathbf{y}^k\}_{k\geq 0}$ be the sequence generated by the FDPG method with $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. Then for any dual optimal solution $\mathbf{y}^*$ of and $k \geq 1$,

$$q_{\mathrm{opt}} - q(\mathbf{y}^k) \leq \frac{2L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{(k+1)^2}.$$

Fast Dual Proximal Gradient - Primal Representation

The Fast Dual Proximal Gradient (FDPG) Method - primal representation

Initialization: $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{V}, t_0 = 1.$
General step $(k \geq 0)$**:**

(a) $\mathbf{u}^k = \underset{\mathbf{u}}{\text{argmax}} \left\{ \langle \mathbf{u}, \mathcal{A}^T(\mathbf{w}^k) \rangle - f(\mathbf{u}) \right\}.$

(b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \frac{1}{L}\mathcal{A}(\mathbf{u}^k) + \frac{1}{L}\mathrm{prox}_{Lg}(\mathcal{A}(\mathbf{u}^k) - L\mathbf{w}^k)$

(c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$

(d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k).$

$O(1/k^2)$ Rate of the Primal Sequence Generated by FDPG

Theorem Let $\{\mathbf{x}^k\}_{k\geq 0}$ and $\{\mathbf{y}^k\}_{k\geq 0}$ be the primal and dual sequences generated by the FDPG method with $L \geq L_F = \frac{\|\mathcal{A}\|^2}{\sigma}$. Then for any optimal dual solution $\mathbf{y}^*$ and $k \geq 1$,

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{4L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{\sigma(k+1)^2}.$$

Proof.

$$\|\mathbf{x}^k - \mathbf{x}^*\|^2 \leq \frac{2}{\sigma}(q_{\rm opt} - q(\mathbf{y}^k)) \leq \frac{2}{\sigma} \cdot \frac{2L\|\mathbf{y}^0 - \mathbf{y}^*\|^2}{(k+1)^2}.$$

$O\left(\frac{1}{k^2}\right)$ rate in function values

[Drusvyatskiy 20']
$O(1/k^2)$ convergence of the primal-dual gap of FISTA.

assumption: $\mathrm{dom}(f^*)$ bounded (implies that f is Lipschitz)

Theorem [primal-dual gap convergence rate]

$$F\left(\sum_{i=0}^{k} \frac{t_i}{t_k^2}\mathbf{x}_i\right) - q(\mathbf{y}^k) \leq \frac{2R^2}{\sigma(k+2)^2}$$

ergodic rate of convergence.

Example 1: Orthogonal Projection onto a Polyhedral set

$$(P_1) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2 : \mathbf{A}\mathbf{x} \le \mathbf{b} \right\}.$$

- Fits model (P) with $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^p$, $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2$,

$$g(\mathbf{z}) = \delta_{\mathrm{Box}[-\infty\mathbf{e},\mathbf{b}]}(\mathbf{z}) = \begin{cases} \mathbf{0}, & \mathbf{z} \le \mathbf{b}, \\ \infty, & \text{else.} \end{cases}$$

  and $\mathcal{A}(\mathbf{x}) \equiv \mathbf{A}\mathbf{x}$.
- $\sigma = 1$
- $\operatorname{argmax}_{\mathbf{x}}\{\langle \mathbf{v}, \mathbf{x}\rangle - f(\mathbf{x})\} = \mathbf{v} + \mathbf{d}$ for any $\mathbf{v} \in \mathbb{R}^n$;
- $\|\mathcal{A}\| = \|\mathbf{A}\|_{2,2}$;
- $\mathcal{A}^T(\mathbf{y}) = \mathbf{A}^T\mathbf{y}$ for any $\mathbf{y} \in \mathbb{R}^p$;
- $\mathrm{prox}_{Lg}(\mathbf{z}) = P_{\mathrm{Box}[-\infty\mathbf{e},\mathbf{b}]}(\mathbf{z}) = \min\{\mathbf{z}, \mathbf{b}\}$.

DPG and FDPG for solving (P_1)

```
**Algorithm 1** [DPG for solving $(P_1)$]
  ▶ **Initialization:** $L \geq \|\mathbf{A}\|_{2,2}^2, \mathbf{y}^0 \in \mathbb{R}^p$.
  ▶ **General Step** $(k \geq 0)$:
      (a) $\mathbf{x}^k = \mathbf{A}^T \mathbf{y}^k + \mathbf{d}$;
      (b) $\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{L}\mathbf{A}\mathbf{x}^k + \frac{1}{L}\min\{\mathbf{A}\mathbf{x}^k - L\mathbf{y}^k, \mathbf{b}\}$.
```

```
**Algorithm 2** [FDPG for solving $(P_1)$]
  ▶ **Initialization:** $L \geq \|\mathbf{A}\|_{2,2}^2, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{R}^p, t_0 = 1$.
  ▶ **General Step** $(k \geq 0)$:
      (a) $\mathbf{u}^k = \mathbf{A}^T \mathbf{w}^k + \mathbf{d}$;
      (b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \frac{1}{L}\mathbf{A}\mathbf{u}^k + \frac{1}{L}\min\{\mathbf{A}\mathbf{u}^k - L\mathbf{w}^k, \mathbf{b}\}$;
      (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
      (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.
```

Example $1\frac{1}{2}$: Orthogonal Projection onto the Intersection of Closed Convex Sets

$$(P_2) \quad \min_{\mathbf{x}\in\mathbb{E}} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2 : \mathbf{x} \in \cap_{i=1}^p C_i \right\}.$$

- $C_1, C_2, \ldots, C_p \subseteq \mathbb{E}$ closed and convex.
- $\mathbf{d} \in \mathbb{E}$.
- Assume that $\cap_{i=1}^p C_i \neq \emptyset$ and that projecting onto each set C_i is an easy task.
- (P_2) fits model (P) with $\mathbb{V} = \mathbb{E}^p, f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2, g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = \sum_{i=1}^p \delta_{C_i}(\mathbf{x}_i)$ and $\mathcal{A} : \mathbb{E} \to \mathbb{V}, \mathcal{A}(\mathbf{z}) = (\underbrace{\mathbf{z}, \mathbf{z}, \ldots, \mathbf{z}}_{p \text{ times}})$
- $\operatorname*{argmax}_{\mathbf{x}}\{\langle \mathbf{v}, \mathbf{x}\rangle - f(\mathbf{x})\} = \mathbf{v} + \mathbf{d}$ for any $\mathbf{v} \in \mathbb{E}$;
- $\|\mathcal{A}\|^2 = p$;
- $\sigma = 1$;
- $\mathcal{A}^T(\mathbf{y}) = \sum_{i=1}^p y_i$ for any $\mathbf{y} \in \mathbb{E}^p$;
- $\mathrm{prox}_{Lg}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) = (P_{C_1}(\mathbf{v}_1), P_{C_2}(\mathbf{v}_2), \ldots, P_{C_p}(\mathbf{v}_p))$ for any $\mathbf{v} \in \mathbb{E}^p$.

DPG and FDPG for Solving (P_2)

Algorithm 3 [DPG for solving (P_2)]

- **Initialization:** $L \geq p, \mathbf{y}^0 \in \mathbb{E}^p$.
- **General Step ($k \geq 0$):**

  (a) $\mathbf{x}^k = \sum_{i=1}^p \mathbf{y}_i^k + \mathbf{d}$;
  (b) $\mathbf{y}_i^{k+1} = \mathbf{y}_i^k - \frac{1}{L}\mathbf{x}^k + \frac{1}{L}P_{C_i}(\mathbf{x}^k - L\mathbf{y}_i^k), \ i = 1, 2, \ldots, p.$

Algorithm 4 [FDPG for solving (P_2)]

- **Initialization:** $L \geq p, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}^p, t_0 = 1.$
- **General Step ($k \geq 0$):**

  (a) $\mathbf{u}^k = \sum_{i=1}^p \mathbf{w}_i^k + \mathbf{d}$;
  (b) $\mathbf{y}_i^{k+1} = \mathbf{w}_i^k - \frac{1}{L}\mathbf{u}^k + \frac{1}{L}P_{C_i}(\mathbf{u}^k - L\mathbf{w}_i^k)$,
  $i = 1, 2, \ldots, p$;
  (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
  (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k).$

Orthogonal Projection onto a Polyhedral Set Revisited

- Algorithm 4 can also be used to find an orthogonal projection of a point $\mathbf{d} \in \mathbb{R}^n$ onto the polyhedral set $C = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$, where $\mathbf{A} \in \mathbb{R}^{p \times n}, \mathbf{b} \in \mathbb{R}^p$.
- C can be written as $C = \cap_{i=1}^p C_i$, where $C_i = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}_i^T \mathbf{x} \leq b_i\}$ with $\mathbf{a}_1^T, \mathbf{a}_2^T, \ldots, \mathbf{a}_p^T$ being the rows of $\mathbf{A}$.
- $P_{C_i}(\mathbf{x}) = \mathbf{x} - \frac{[\mathbf{a}_i^T \mathbf{x} - b_i]_+}{\|\mathbf{a}_i\|^2} \mathbf{a}_i$.

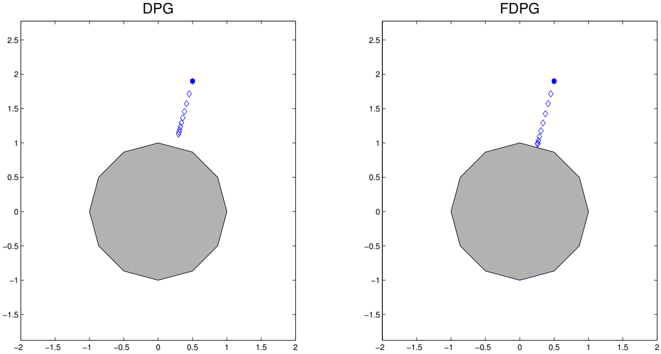
Algorithm 5 [FDPG for solving (P_1)]

- **Initialization:** $L \geq p, \mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}^p, t_0 = 1$.
- **General Step ($k \geq 0$):**

```
(a) $\mathbf{u}^k = \sum_{i=1}^p \mathbf{w}_i^k + \mathbf{d}$;
(b) $\mathbf{y}_i^{k+1} = -\frac{1}{L\|\mathbf{a}_i\|^2}[\mathbf{a}_i^T(\mathbf{u}^k - L\mathbf{w}_i^k) - b_i]_+ \mathbf{a}_i, \ i = 1, 2, \ldots, p$;
(c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
(d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.
```

Comparison Between DPG and FDPG – Numerical Example

- Consider the problem of projecting the point $(0.5, 1.9)^T$ onto a dodecagon - a regular polygon with 12 edges represented as the intersection of 12 half-spaces.
- The first 10 iterations of the DPG (Algorithm 3) and FDPG (Algorithm 4/5) methods with $L = p = 12$ can be seen below.



Example 2: One-Dimensional Total Variation Denoising

$$(P_3) \quad \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2 + \lambda\|\mathbf{D}\mathbf{x}\|_1 \right\},$$

- Fits model (P) with
  $\mathbb{E} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^{n-1}, f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|^2 (\sigma = 1), g(\mathbf{y}) = \lambda\|\mathbf{y}\|_1$, $\mathcal{A}(\mathbf{x}) \equiv \mathbf{D}\mathbf{x}$
- $\underset{\mathbf{x}}{\operatorname{argmax}}\{\langle \mathbf{v}, \mathbf{x}\rangle - f(\mathbf{x})\} = \mathbf{v} + \mathbf{d}$ for any $\mathbf{v} \in \mathbb{E}$;
- $\|\mathcal{A}\|^2 = \|\mathbf{D}\|_{2,2}^2 \le 4$;
- $\sigma = 1$;
- $\mathcal{A}^T(\mathbf{y}) = \mathbf{D}^T\mathbf{y}$ for any $\mathbf{y} \in \mathbb{R}^{n-1}$;
- $\mathrm{prox}_{Lg}(\mathbf{y}) = \mathcal{T}_{\lambda L}(\mathbf{y})$.

Example 3 Contd.

Algorithm 6 [DPG for solving (P_3)]

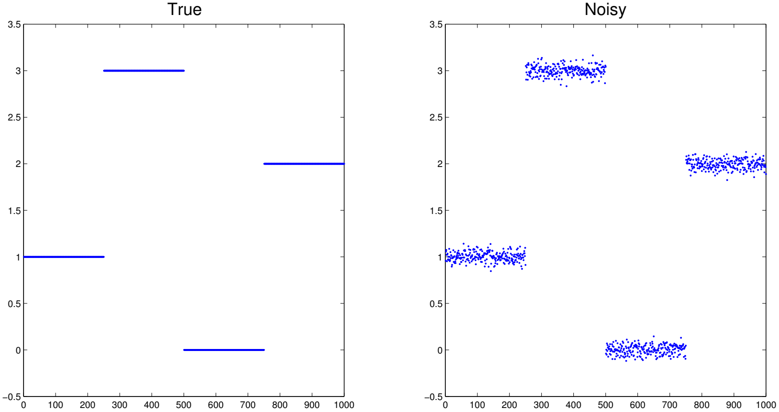
```
▶ Initialization: $\mathbf{y}^0 \in \mathbb{R}^{n-1}$.
▶ General Step ($k \geq 0$):
    (a) $\mathbf{x}^k = \mathbf{D}^T\mathbf{y}^k + \mathbf{d}$;
    (b) $\mathbf{y}^{k+1} = \mathbf{y}^k - \frac{1}{4}\mathbf{D}\mathbf{x}^k + \frac{1}{4}\mathcal{T}_{4\lambda}(\mathbf{D}\mathbf{x}^k - 4\mathbf{y}^k)$.
```

Algorithm 7 [FDPG for solving (P_3)]

```
▶ Initialization: $\mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{R}^{n-1}$, $t_0 = 1$.
▶ General Step ($k \geq 0$):
    (a) $\mathbf{u}^k = \mathbf{D}^T\mathbf{w}^k + \mathbf{d}$;
    (b) $\mathbf{y}^{k+1} = \mathbf{w}^k - \frac{1}{4}\mathbf{D}\mathbf{u}^k + \frac{1}{4}\mathcal{T}_{4\lambda}(\mathbf{D}\mathbf{u}^k - 4\mathbf{w}^k)$;
    (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
    (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.
```

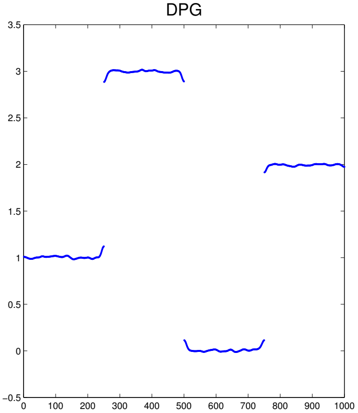
Numerical Example

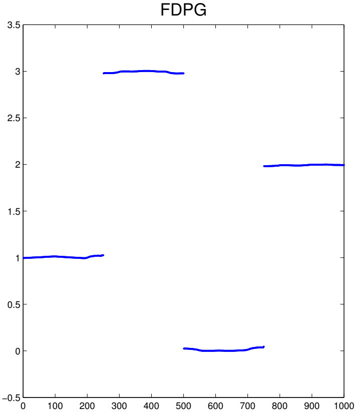
- $n = 1000$
- $\mathbf{d}$ is a noisy measurement of a step function.



Numerical Example Contd.

- 100 iterations of Algorithms 6 (DPG) and 7 (FDPG) initialized with $\mathbf{y}^0 = \mathbf{0}$.





- Objective function values of the DPG and FDPG methods after 100 iterations are 9.1667 and 8.4621 respectively; the optimal value is 8.3031.

The Dual Block Proximal Gradient Method

The Model

$$(Q) \quad \min_{\mathbf{x}\in\mathbb{E}} \left\{ f(\mathbf{x}) + \textstyle\sum_{i=1}^{p} g_i(\mathbf{x}) \right\}.$$

Underlying Assumptions.

(A) $f : \mathbb{E} \to (-\infty, +\infty]$ is proper closed and σ-strongly convex ($\sigma > 0$).

(B) $g_i : \mathbb{E} \to (-\infty, +\infty]$ is proper closed and convex for any $i \in \{1, 2, \ldots, p\}$.

(C) $\mathrm{ri}(\mathrm{dom}(f)) \cap (\cap_{i=1}^{p} \mathrm{ri}(\mathrm{dom}(g_i))) \neq \emptyset$.

Problem (Q) fits model (P) with
$\mathbb{V} = \mathbb{E}^p, g(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p) = \sum_{i=1}^{p} g_i(\mathbf{x}_i), \mathcal{A}(\mathbf{z}) = (\underbrace{\mathbf{z}, \mathbf{z}, \ldots, \mathbf{z}}_{p \text{ times}})$.

- $\|\mathcal{A}\|^2 = p$;
- $\mathcal{A}^T(\mathbf{y}) = \sum_{i=1}^{p} y_i$ for any $\mathbf{y} \in \mathbb{E}^p$;
- $\mathrm{prox}_{Lg}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p) = (\mathrm{prox}_{Lg_1}(\mathbf{v}_1), \mathrm{prox}_{Lg_2}(\mathbf{v}_2), \ldots, \mathrm{prox}_{Lg_p}(\mathbf{v}_p))$

FDPG for Solving (Q)

```
Algorithm 9 [FDPG for solving (Q)]
  ▶ Initialization: $\mathbf{w}^0 = \mathbf{y}^0 \in \mathbb{E}^p$, $t_0 = 1$.
  ▶ General Step ($k \geq 0$):
      (a) $\mathbf{u}^k = \operatorname{argmax}_{\mathbf{u}\in\mathbb{E}} \left\{ \left\langle \mathbf{u}, \sum_{i=1}^{p} \mathbf{w}_i^k \right\rangle - f(\mathbf{u}) \right\}$;
      (b) $\mathbf{y}_i^{k+1} = \mathbf{w}_i^k - \frac{\sigma}{p}\mathbf{u}^k + \frac{\sigma}{p}\mathrm{prox}_{\frac{p}{\sigma}g_i}(\mathbf{u}^k - \frac{p}{\sigma}\mathbf{w}_i^k)$, $i = 1, 2, \ldots, p$;
      (c) $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$;
      (d) $\mathbf{w}^{k+1} = \mathbf{y}^{k+1} + \left(\frac{t_k - 1}{t_{k+1}}\right)(\mathbf{y}^{k+1} - \mathbf{y}^k)$.
```

Exercise 2: One-Dimensional Total Variation Denoising

Consider the problem

$$(T) \quad f_{\rm opt} = \min_{\mathbf{x}\in\mathbb{R}^n} \left\{ F(\mathbf{x}) \equiv \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2 + \lambda \sum_{i=1}^{n-1} |x_{i-1} - x_i| \right\},$$

where $\mathbf{d} \in \mathbb{R}^n$ and $\lambda > 0$.

- The problem can be written as $\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2 + \lambda\|\mathbf{D}\mathbf{x}\|_1$. What is $\mathbf{D}$?
- Write the FDPG method for solving (T), including all the constants.
- (T) can be written as $\min_{\mathbf{x}\in\mathbb{R}^n}\{f(\mathbf{x}) + g_1(\mathbf{x}) + g_2(\mathbf{x})\}$, where

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}-\mathbf{d}\|_2^2, g_1(\mathbf{x}) = \lambda \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} |x_{2i-1} - x_{2i}|, g_2(\mathbf{x}) = \lambda \sum_{i=1}^{\lfloor \frac{n-1}{2} \rfloor} |x_{2i} - x_{2i+1}|.$$

- Write explicitly the prox operators of g_1 and g_2. Write the FDPG method method for solving (T) using the decomposition $g_1 + g_2$.

Exercise 3: Soft Margin SVM

Given a set of data points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$ and corresponding labels $y_1, y_2, \dots, y_n$. The soft margin SVM problem is given by

$$\min \left\{ \frac{1}{2}\|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \max\{0, 1 - y_i \mathbf{w}^T \mathbf{x}_i\} \right\}.$$

(a) Write explicitly (including e.g. prox computations, Lipschitz constants) the DPG and FDPG for solving the problem.

(b) Suppose that the data is generated as follows:

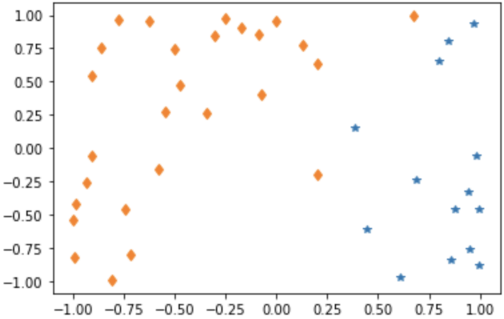
MATLAB

```
x=sin(10*[1:40].^3)’;
y=sin(28*[1:40].^3)’;
cl=[2*x<y+0.5]+1;
x = [x;0.2];
y = [y;-0.2];
cl = [cl;2];
A1=[x(cl==1),y(cl==1)];
A2=[x(cl==2),y(cl==2)];
figure(1)
plot(A1(:,1),A1(:,2),’*’,’MarkerSize’,6)
hold on
plot(A2(:,1),A2(:,2),’d’,’MarkerSize’,6)
```

Python

```
import numpy as np
import matplotlib.pyplot as plt
x = np.sin(10*np.arange(1,41)**3)
y = np.sin(28*np.arange(1,41)**3)
cl = (2*x<y+0.5)+1
x = np.hstack((x,0.2))
y = np.hstack((y,-0.2))
cl = np.hstack((cl,2))
A1=np.column_stack((x[cl==1], y[cl==1]))
A2=np.column_stack((x[cl==2], y[cl==2]))
plt.plot(A1[:,0], A1[:,1],’*’)
plt.plot(A2[:,0], A2[:,1],’d’)
plt.show()
```

Exercise 3 Contd.



Write a MATLAB/Python code that implements 40 iterations of the DPG and FDPG methods. Write the solutions produced by each of the methods. For each of the solutions, plot the two classes of points along with the corresponding hyperplane.

References

- A. Beck, *First Order Methods in Optimization* (2017).
- A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. (2009).
- A. Beck and M. Teboulle, *A fast dual proximal gradient algorithm for convex minimization and applications*, Oper. Res. Lett. (2014)
- A. Beck and M. Teboulle. *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*, IEEE Trans. Image Process. (2009)
- H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces* (2011).
- A. Chambolle, *An algorithm for total variation minimization and applications*, J. Math. Imaging Vision (2004)
- P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward backward splitting*, Multiscale Model. Simul. (2005).
- N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends in Optim. (2014).
- Y. Nesterov, *Gradient methods for minimizing composite fun.*, Math. Program. (2013)
- P. Tseng, *Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. SIAM J. Control Optim.*, (1991)