Mini-course on Sparse estimation off-the-grid

Algorithms

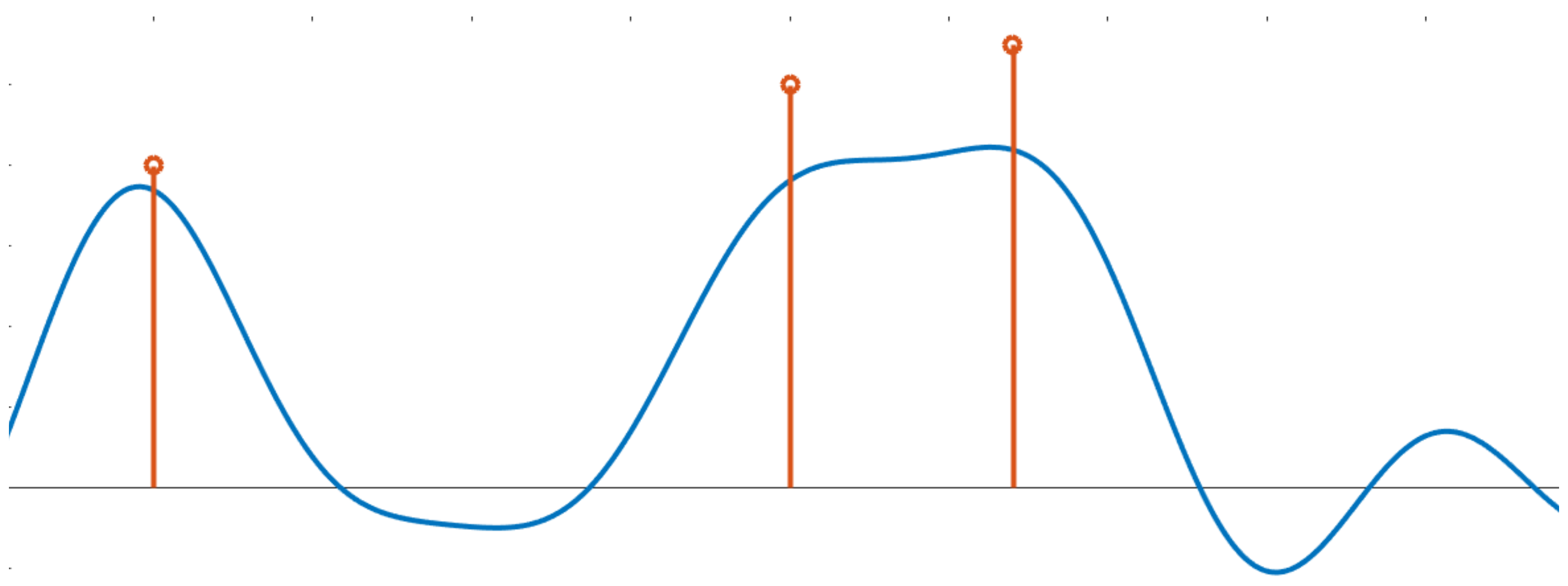
Clarice Poon

Discretise on a fine grid?

Approach: Discrete Φ on a **fine grid**

For $\phi(x) \in \mathbb{R}^m$, define: $A = [\phi(x_1), \phi(x_2), \cdots, \phi(x_N)] \in \mathbb{R}^{m \times N}$

$$\min_{x \in \mathbb{R}^n} F(x) = \frac{1}{2\lambda} \|Ax - y\|^2 + \|x\|_1$$



Example: Fourier measurements

Column i : $A_i = \left(\exp(2\pi\sqrt{-1} x_i^\top \omega_k) \right)_k$

$\mathcal{O}(p^{-d})$ grid points if $[x_i] \subseteq [0,1]^d$ spaced p apart

Forward-Backward splitting

$$\min_x F(x) := f(x) + g(x)$$

Assume that f is differentiable

Forward-Backward splitting:

$$\begin{cases} \hat{x}_{k+1} & = x_k - \tau \nabla f(x_k) \\ x_{k+1} & = \mathrm{Prox}_{\tau g}(\hat{x}_{k+1}) := \mathrm{argmin}_z \frac{1}{2\tau} \|z - \hat{x}_{k+1}\|^2 + g(z) \end{cases}$$

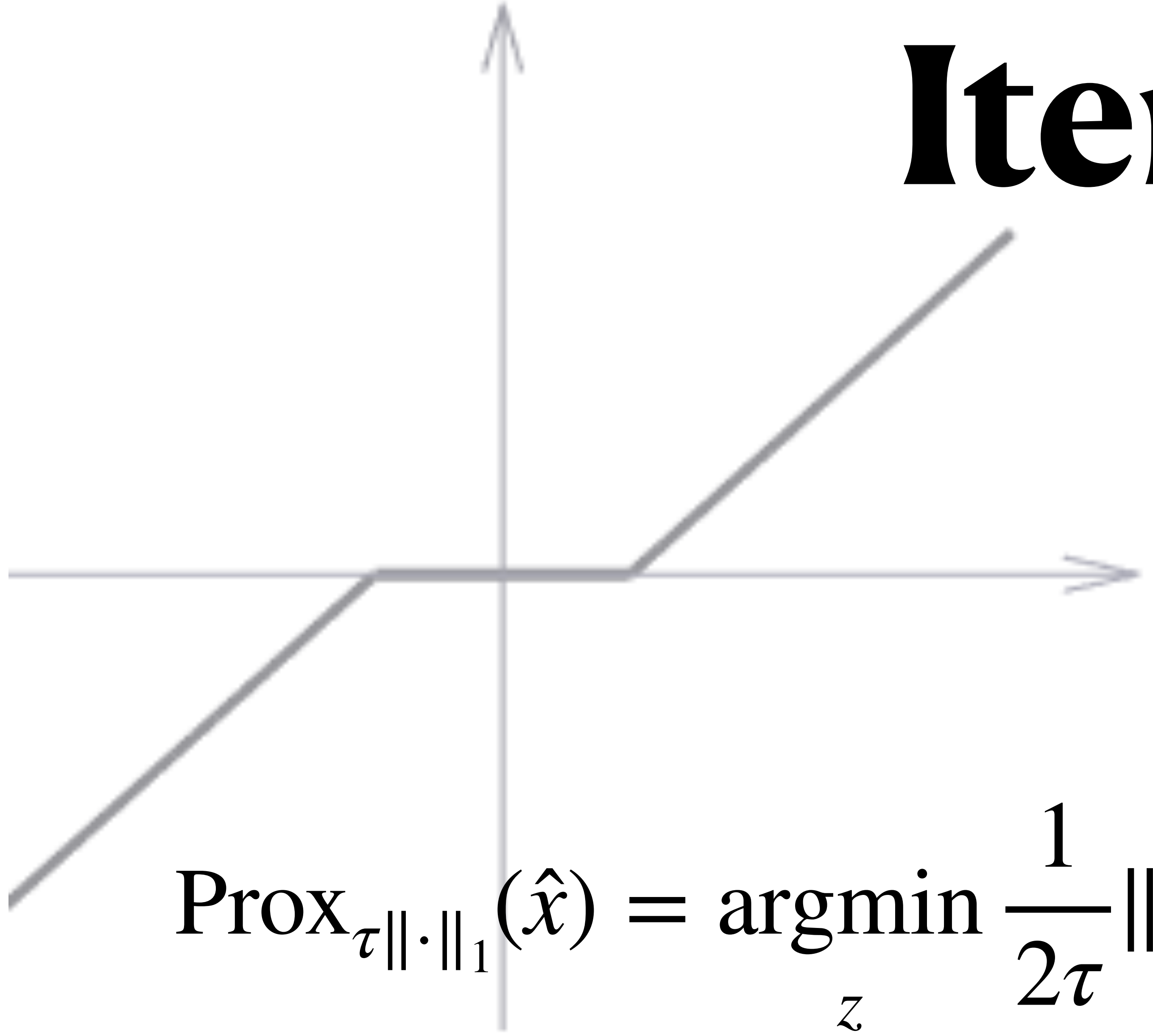
Assume:

- f, g convex
- $\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|$

Convergence rates: If $\tau = 1/L$, then

$$F(x_k) - \min_x F(x) \leq \frac{L}{k}$$

Iterative Soft Thresholding



$$\begin{cases} \hat{x}_{k+1} & = x_k - \tau A^\top (A x_k - y) \\ x_{k+1} & = \mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}_{k+1}) \end{cases}$$

$$\begin{aligned} \mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}) &= \underset{z}{\mathrm{argmin}} \frac{1}{2\tau}\|z - \hat{x}\|^2 + \|z\|_1 \\ &= \mathrm{sign}(\hat{x})(|\hat{x}| - \tau)_+ \end{aligned}$$

$f(x) = \dfrac{1}{2}\|Ax - y\|^2$ is L-Lipschitz with $L = \|A\|^2$

Convergence rates: for $\tau = 1/L$

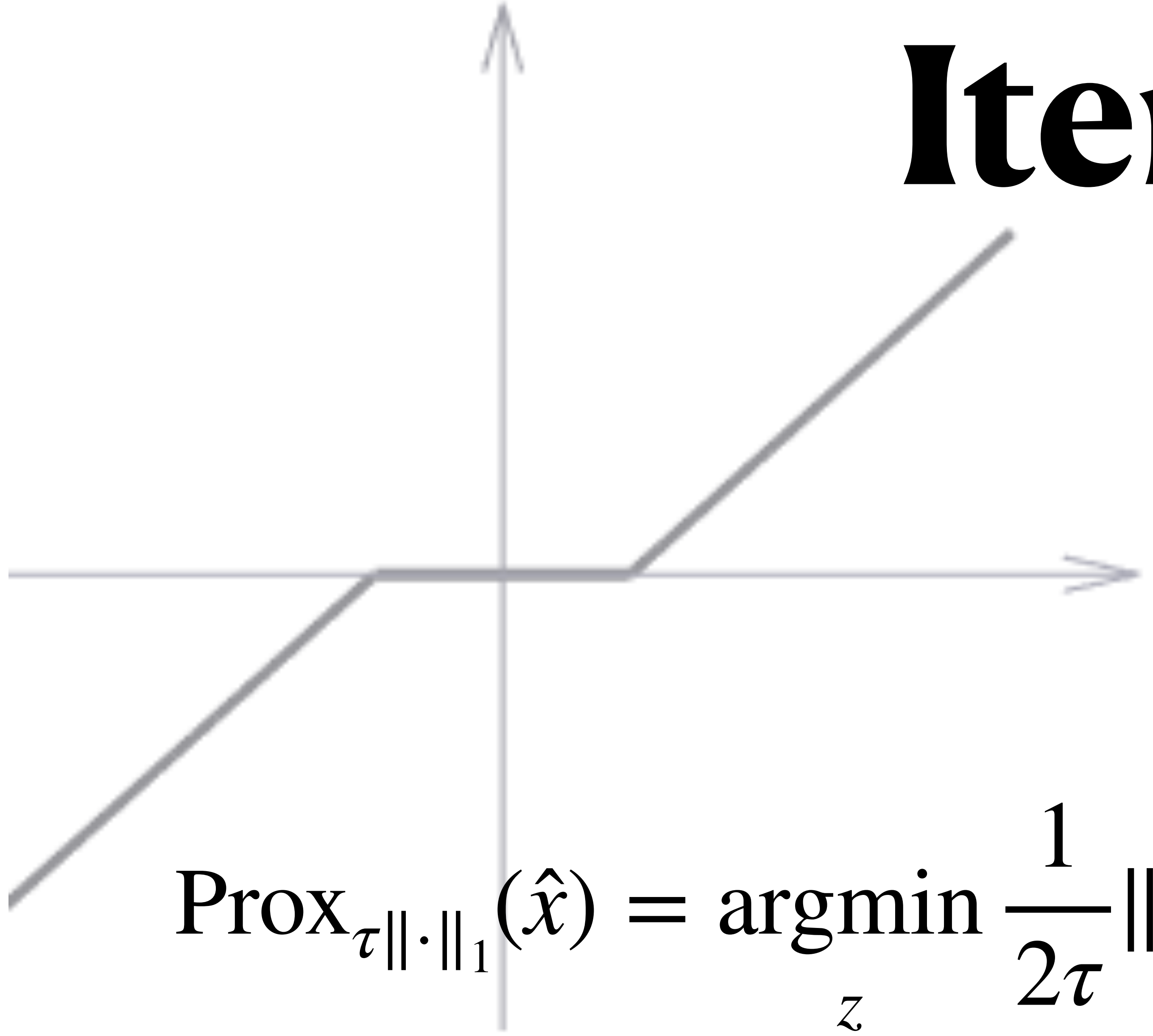
$$F(x_k) - \min_x F(x) \leq \frac{\|A\|^2}{k}$$

Example: Fourier measurements

Column i : $A_i = \left(\exp(2\pi\sqrt{-1} x_i^\top \omega_k) \right)_k$

$$\|A\|^2 = \mathcal{O}(p^{-d})$$

Iterative Soft Thresholding



$$\min_{x \in \mathbb{R}^n} \Phi(x) = \frac{1}{2\lambda}\|Ax - y\|^2 + \|x\|_1$$

$$\begin{aligned} \mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}) &= \operatorname*{argmin}_z \frac{1}{2\tau}\|z - \hat{x}\|^2 + \|z\|_1 \\ &= \mathrm{sign}(\hat{x})(|\hat{x}| - \tau)_+ \end{aligned}$$

$$\begin{cases} \hat{x}_{k+1} &= x_k - \tau A^\top(Ax_k - y) \\ x_{k+1} &= \mathrm{Prox}_{\tau\|\cdot\|_1}(\hat{x}_{k+1}) \end{cases}$$

Convergence rates (depends on n):

$$F(x_k) - \min_x F(x) \leq \frac{C_n}{k}$$

NB: C_n can grow with n!

Grid-free convergence rates (Chizat 2021):

$$F(x_k) - \min_x F(x) \leq k^{-2/(d+1)}$$

NB: Result is independent of n

Issue with fine grids: exact resolution is not possible [Duval and Peyré 2017]

Finite dimensional formulations

$$\sup_{p \in \mathbb{R}^m} \langle p, y \rangle - \frac{\lambda}{2} \|p\|^2 \quad \text{s.t.} \quad \|\Phi^* p\|_\infty \leq 1$$

In general, the constraint is infinite dimensional. However, there are special cases for which one can formulate as a finite dimensional problem.

Fourier setting
[Candés & Fernandez-Granda 2014]:

Minimise over all p such that for all x,
$$|\sum_{|k| \leq f_c} p_k \exp(2\pi\sqrt{-1}kx)| \leq 1$$
This constraint can be written as a positive semidefinite constraint on matrices.

Quadratic: $\phi(x) = ((u_i^\top x)^2)_i$ for $x \in \mathbb{S}_{n-1}$, then
$$\Phi^* p(x) = \sum_k p_k (u_k^\top x)^2 = \langle (\sum_k p_k u_k u_k^\top) x, x \rangle$$
Constraint: Spectral norm is bounded by 1.

ReLU [Pilanci & Ergen 2020]: $\phi(x) = ((u_i^\top x)_+)_i$ for $x \in \mathbb{S}_{n-1}$. Then. $\Phi^* p(x) = \sum_k p_k (u_k^\top x)_+$

Uses the fact that there are only finitely many support patterns for which (Ux)_+

Semi-definite programming

Special case of Fourier samples:

minimise over all p such that for all x, $\left| \sum_{|k| \leq f_c} p_k \exp(2\pi\sqrt{-1}kx) \right| \leq 1$

Theorem (Dumitrescu): A trigonometric polynomial $f(t) = \sum_{k=0}^{n-1} c_k \exp(\sqrt{-1}2\pi kx)$ with $p \in \mathbb{C}^n$ lid uniformly bounded by 1 in magnitude if there exists $Q \in \mathbb{C}^{n\times n}$ Hermitian s.t.

$$0 \preceq \begin{pmatrix} Q & p \\ p^* & 1 \end{pmatrix} \quad \text{and} \quad \sum_{i=1}^{n-j} Q_{i,i+j} = \delta_{0,j}$$

Semi-definite programming

Equivalent dual formulation: Let $n = 2f_c + 1$:

$$\sup_{p,Q} \langle p, y \rangle - \frac{\lambda}{2} \|p\|^2 \quad \text{s.t.} \quad 0 \preceq \begin{pmatrix} Q & p \\ p^* & 1 \end{pmatrix} \quad \text{and} \quad \sum_{i=1}^{n-j} Q_{i,i+j} = \delta_{0,j}$$

Finite dimensional semi-definite program!

1. Solve SDP to find p_λ

2.Find the support of m_λ by finding the roots of the polynomial

$$f_{2n-2}(e^{\sqrt{-1}2\pi x}) = 1 - |\Phi^* p_\lambda(x)|^2$$

3.This has at most $n-1$ roots (unless identically 0)

4.Solve for amplitudes.

Frank-Wolfe algorithm

$$\min_{x \in C} f(x)$$

C is a weakly compact convex set of a Banach space.

f is a differentiable convex function.

1. $z^k \in \operatorname{argmin}_{z \in C} f(x^k) + \langle \nabla f(x^k), z - x^k \rangle$
2. If $\langle \nabla f(x^k), z^k - x^k \rangle = 0$ then x^k is a solution.
3. $\gamma^k = 2/(k+2)$
4. $x^{k+1} = x^k + \gamma^k (z^k - x^k)$

- f is convex $\Longrightarrow f(z) \geq f(x^k) + \langle \nabla f(x^k), z - x^k \rangle \overset{Step2}{\Longrightarrow} f(z) \geq f(x^k)$ for all z.
- One can replace x^{k+1} in step 4 with any $\hat{x}^{k+1}$ such that $f(\hat{x}^{k+1}) \leq f(x^{k+1})$.

Applying Frank-Wolfe to the Blasso

$$\mu \in \underset{\mu \in \mathscr{M}(\mathscr{X})}{\operatorname{argmin}} f_\lambda(m) := \frac{1}{2}\|\Phi\mu - y\|^2 + \lambda\|\mu\|_{TV}$$

$\mu \in \operatorname{argmin} f_\lambda(\mu)$ implies

$$\lambda\|\mu\|_{TV} \leq \lambda\|\mu\|_{TV} + \frac{1}{2}\|\Phi\mu - y\|^2$$
$$\leq \lambda\|0\|_{TV} + \frac{1}{2}\|\Phi 0 - y\|^2 = \frac{1}{2}\|y\|^2$$

$$(\|\mu\|_{TV}, \mu) \in \underset{(t,\mu) \in C}{\operatorname{argmin}} \hat{f}_\lambda(t, \mu) := \frac{1}{2}\|\Phi\mu - y\|^2 + \lambda t$$

$$C = \{(t, \mu) \in \mathbb{R}_+ \times \mathscr{M}(\mathscr{X}) \ : \ \|\mu\|_{TV} \leq t \leq \|y\|^2/(2\lambda) =: M \}$$

$\hat{f}_\lambda$ is differentiable :
$$\begin{cases} \partial_t f = \lambda \\ \partial_\mu \hat{f}_\lambda(t, \mu) = \Phi^*(\Phi\mu - y) \end{cases}$$

[Bredies and Pikkarainnen 2013, Boyd et al, 2017, Denoyelle et al 2018]

Convergence of Frank-Wolfe

[Jaggi (2011)]: The curvature constant of f over C is

$$R := \max_{\substack{\gamma \in [0,1] \\ x, s, y \in C \\ y = (1-\gamma)x + \gamma s}} \frac{2}{\gamma^2}\left(f(y) - f(x) - \langle \nabla f(x), y - x \rangle\right)$$

Convergence rate:

$$f(x^k) - f^* \leq \frac{2R}{k+2}$$

- $R = 0$ if f is linear
- ∇f is L-Lipschitz $\implies \quad f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \dfrac{L}{2}\|y - x\|^2$

  $\implies \quad R \leq L\mathrm{diam}(C)^2$
- For $\hat{f}_\lambda$, $R = \dfrac{1}{2}\sup\{\|\Phi(m - m')\|^2 : \|m\|_{TV}, \|m'\|_{TV} \leq \|y\|^2/2\}$
- If $\|\phi(x)\| = 1$ for all x, then $R \lesssim \|y\|^2$ and the convergence rate is $\mathcal{O}(\|y\|^2/k)$

Applying Frank-Wolfe to the Blasso

Key step. Let $u^k = (t^k, m^k)$.

$$z^k \in \operatorname{argmin}_{z\in C} \hat{f}(u^k) + \langle \nabla \hat{f}(u^k), z - u^k \rangle$$

$$= \operatorname{argmin}_{(t,m)\in C} \langle \Phi^*(\Phi m^k - y), m \rangle + \langle \lambda, t \rangle$$

C is a convex set and minimum is achieved at an extremal point of C.

$$E = \{(M, \pm M\delta_x) \ : \ x \in \mathcal{X}\}$$

$$\implies \quad z^k = (M, \pm M\delta_{x_k})$$

$$x^k \in \operatorname*{argmin}_{x\in\mathcal{X}} \pm [\Phi^*(\Phi m^k - y)](x) + \lambda$$

$$= \operatorname*{argmax}_{x\in\mathcal{X}} \ |\eta^k(x)|, \qquad \eta^k = \frac{1}{\lambda}[\Phi^*(\Phi m^k - y)]$$

$$a^k = -\operatorname{sign}(\eta^k(x^k))\, M$$

Applying Frank-Wolfe to the Blasso

At each iteration k, $\mu^{k-1} = \sum_{j=1}^{k-1} a^j \delta_{x^j}$

Define $\eta^k = \frac{1}{\lambda}[\Phi^*(\Phi\mu^{k-1} - y)]$ and $\gamma^k = 2/(k+2)$.

1. Add new spike: $x^k = \mathrm{argmax}_{x\in\mathscr{X}} |\eta^k(x)|$
2. $\mu^k = \mu^{k-1} + \gamma^k \left(a_k \delta_{x^k} - \mu^{k-1}\right)$

Terminate if $\eta^k(x^k) = \pm 1$ and return μ^k

Sliding Frank-Wolfe: *[Denoyelle et al 2018]*

Replace step 2 with any $\hat{\mu}^k$ such that $f_\lambda(\hat{\mu}^k) \leq f_\lambda(\mu^k)$:

$\hat{\mu}^k = \sum_{j=1}^{k} \hat{a}_j \, \delta_{\hat{x}_j}$ where

$$(\hat{a}, \hat{x}) \in \min_{a,x} \lambda \|a\|_1 + \frac{1}{2}\|\Phi\mu_{a,x} - y\|^2$$

Theorem (Denoyelle et al, 2019): finite convergence if η_V is non-degenerate at the solution.

Remarks

- The sliding Frank-Wolfe is an off-the-grid algorithm, however:

  ➡ the **difficulty** is in the step $x^k = \text{argmax}_{x \in \mathcal{X}} |\eta^k(x)|$

- $\mathcal{X}$ is a continuous space and η^k is a continuous (smooth) function
- In practice, discretize $\mathcal{X}$ and do a local ascent step on η^k
- This is computationally intensive if $\mathcal{X} \subset \mathbb{R}^d$ and d is large.

Particle methods

$$\min_{\mu} \lambda \|\mu\|_{TV} + \frac{1}{2}\|\Phi\mu - y\|^2$$

This has a solution consisting of $m + 1$ Diracs

$$\min_{a,x} \lambda \sum_{i=1}^{k} |a_i| + \frac{1}{2}\| \sum_{i=1}^{k} \phi(x_i)a_i - y\|^2$$

Has the same value as $P_\lambda(y)$ when $k \geq m$

Chizat and Bach (2018): Global convergence results for *sufficiently large k*.

VarPro

[Golub and Pereyra, 1978]

$$\min_{a,x} \| \sum_{j=1}^{K} \phi(x_j) a_j - y \|^2$$

$\Longleftrightarrow$

$$\min_{x} f(x) \quad \text{where}$$

$$f(x) = \min_{a} \| \sum_{j} \phi(x_j) a_j - y \|^2$$

Easy to compute gradient:

$$\partial_{x_j} f(x) = \partial_{x_j} \| \Phi_x \bar{a} - y \|^2 = \bar{a}_j \nabla \phi(x_j)^\top (\Phi_x \bar{a} - y)$$

$$\bar{a} = \mathrm{argmin}_a \| \Phi_x a - y \|^2 = \Phi_x^\dagger y$$

In general, solution is non-sparse, except in special cases, e.g. ReLU

Leads to better problem conditioning.

VarPro

A practical approach for sparse solutions (Nonsmooth VarPro):

$$\min_x f(x) \quad \text{where}\, f(x) := \min_a \frac{1}{2}\|\Phi_x a - y\|^2 + \lambda\|a\|_1$$

If the inner Lasso problem has a unique solution for x, then f is differentiable at x with gradient

$$\partial_{x_j} f(x) = \partial_{x_j}\|\Phi_x \bar{a} - y\|^2 = \bar{a}_j \nabla\phi(x_j)^\top(\Phi_x \bar{a} - y)$$

$$\bar{a} = \operatorname{argmin}_a \frac{1}{2}\|\Phi_x a - y\|^2 + \lambda\|a\|_1$$

Summary

- For certain settings (Fourier sampling), one can convert to a finite dimensional optimisation problem.
- The Frank-Wolfe algorithm is a versatile algorithm for computations off-the-grid. Works well in low dimensions, but there is a difficulty with finding the argmax of η_k
- Particle methods are effective in practice, but no quantitative rates.

References

SDP/convex finite dimensional formulations

- Candès, E. J., & Fernandez-Granda, C. (2014). Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, *67*(6), 906-956.
- Catala, Paul, Vincent Duval, and Gabriel Peyré. "A low-rank approach to off-the-grid sparse superresolution." *SIAM Journal on Imaging Sciences* 12.3 (2019): 1464-1500.
- Pilanci, Mert, and Tolga Ergen. "Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks." *International Conference on Machine Learning*. PMLR, 2020.

Frank-Wolfe

- Bredies, K., & Pikkarainen, H. K. (2013). Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, *19*(1), 190-218.
- Boyd, N., Schiebinger, G., & Recht, B. (2017). The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, *27*(2), 616-639.
- Denoyelle, Q., Duval, V., Peyré, G., & Soubies, E. (2019). The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, *36*(1), 014001.

“Particle” approaches

- Chizat, L., & Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, *31*.
- Golub, Gene, and Victor Pereyra. "Separable nonlinear least squares: the variable projection method and its applications." *Inverse problems* 19.2 (2003): R1.