Supplementary material for "Efficient First-Order Algorithms for Adaptive Signal Denoising", ICML 2018

A. Background

A.1. Adaptive signal denoising

Assume that the goal is to estimate the signal only on $[0, n]$, from observations (1), and consider convolution-type estimators

$$\widehat{x}_t^{\varphi} = [\varphi * y]_t := \sum_{\tau \in \mathbb{Z}} \varphi_\tau y_{t-\tau} \quad 0 \le t \le n. \tag{32}$$

Here, φ is itself an element of $\mathbb{C}(\mathbb{Z})$ called a *filter*; note that if $\varphi \in \mathbb{C}_n(\mathbb{Z})$, (2) defines an estimator of the projection of $x \in \mathbb{C}(\mathbb{Z})$ to $\mathbb{C}_n(\mathbb{Z})$ from observations (1) on $\mathbb{C}_n^{\pm}(\mathbb{Z})$. If the filter φ is fixed and does not depend on the observations, estimator (2) is linear in observations; otherwise it is not. Now, assume, following (Ostrovsky et al., 2016), that $x \in \mathbb{C}(\mathbb{Z})$ belongs to a *shift-invariant* linear subspace $\mathcal{S}$ of $\mathbb{C}(\mathbb{Z})$ – an invariant subspace of the unit shift operator

$$\Delta : \mathbb{C}(\mathbb{Z}) \to \mathbb{C}(\mathbb{Z}), \quad [\Delta x]_t = x_{t-1}.$$

As shown in (Ostrovsky et al., 2016), one can explicitly construct a filter ϕ^o, depending on $\mathcal{S}$, such that the worst-case ℓ_2-risk of the estimator (2) with $\varphi = \phi^o$ satisfies

$$\mathbf{E}^{\frac{1}{2}} \left\{ \|x - \phi^o * y\|_{n,2}^2 \right\} \le \frac{\sigma\rho}{\sqrt{n+1}} \quad \forall x \in \mathcal{S}, \tag{33}$$

where the factor $\rho = \tilde{O}(s^\kappa)$ for some $\kappa > 0$, that is, is polynomial on the subspace dimension $s = \dim(\mathcal{S})$ and logarithmic in the sample size (the logarithmic factor can be dropped in some situations). In fact, one even has a pointwise bound: for any $0 \le \tau \le n$, with prob. $\ge 1 - \delta$,

$$|x_\tau - [\phi^o * y]_\tau| \le \frac{C\sigma\rho\sqrt{1 + \log\left(\frac{n+1}{\delta}\right)}}{\sqrt{n+1}} \quad \forall x \in \mathcal{S}. \tag{34}$$

Note that for any fixed subspace $\mathcal{S}$, not even a shift-invariant one, the worst-case ℓ_2-risk and pointwise risk of *any* estimator can both bounded from below with $c\sqrt{s/(n+1)}$ for some absolute constant c (Johnstone, 2011). Hence, $\widehat{x}^{\phi^o} = \phi^o * y$ is nearly minimax on $\mathcal{S}$ as long as $s \ll n$: its "suboptimality factor" – the ratio of its worst-case ℓ_2-risk to that of a minimax estimator – only depends on the subspace dimension s but not on the sample size n. Unfortunately, $\widehat{x}^{\phi^o}$ depends on subspace $\mathcal{S}$ through the "oracle" filter ϕ^o, and hence it cannot be used in the adaptive estimation setting where the subspace $\mathcal{S}$ with $\dim(\mathcal{S}) = s$ is unknown, but one still would like to attain bounds of the type (33). However, adaptive estimators can be found in the convolution form $\widehat{x} = \widehat{\varphi} * y$ where filter $\widehat{\varphi} = \widehat{\varphi}(y)$ is not fixed anymore, but instead is inferred from the observations. Moreover, $\widehat{\varphi}$ is given as an optimal solution of a certain optimization problem. Several such problems have been proposed, all resting upon a common principle – minimization of the Fourier-domain residual

$$\|F_n[y - \varphi * y]\|_p \tag{35}$$

with regulzarization via the ℓ_1-norm $\|F_n[\varphi]\|_1$ of the DFT of the filter. Such regularization is motivated by the following non-trivial fact, see (Harchaoui et al., 2015b): given an oracle filter $\phi^o \in \mathbb{C}_{\lfloor n/2 \rfloor}(\mathbb{Z})$ which satisfies (33) with n replaced with $3n$, one can point out a new filter $\varphi^o \in \mathbb{C}_n(\mathbb{Z})$ which satisfies a "slightly weaker" counterpart of (34),

$$|x_\tau - [\varphi^o * y]_\tau| \le \frac{3\sigma r\sqrt{1 + \log\left(\frac{n+1}{\delta}\right)}}{\sqrt{n+1}} \quad \forall x \in \mathcal{S} \tag{36}$$

where $r = 2\rho^2$, but also admits a bound on DFT in ℓ_1-norm:

$$\|F_n[\varphi^o]\|_1 \le \frac{r}{\sqrt{n+1}}, \quad r = 2\rho^2. \tag{37}$$

see (Ostrovsky et al., 2016). In fact, (37) is the key property that allows to control the statistical performance of adaptive convolution-type estimators. In some situtaions, polynomial upper bounds on the function $\rho(s)$ are known. Then, adaptive convolution-type estimators with provable statistical guarantees can be obtained by minimizing the residual (4) with $p = \infty$ (Harchaoui et al., 2015b) or $p = 2$ (Ostrovsky et al., 2016) under the constraint (37). A more practical approach is to use penalized estimators, cf. Sec. 1, that attain similar statistical bounds, see (Ostrovsky et al., 2016) and references therein.

A.2. Online accuracy certificates

The guarantees on the accuracy of optimization algorithms presented in Section 2 have a common shortcoming. They are "offline" and worst-case, stated once and for all, for the worst possible problem instance. Neither do they get improved in the course of computation, nor become more optimistic when facing an "easy" problem instance of the class. However, in some situations, online and "opportunistic" bounds on the accuracy are available. Following the terminology introduced in (Nemirovski et al., 2010), such bounds are called *accuracy certificates*. They can be used for early stopping of the algorithm if the goal is to reach some fixed accuracy ε). One situation in which accuracy certificates are available is saddle-point minimization (via a first-order algorithm) in the case where the domains are bounded and admit an efficiently computable *linear maximization oracle*. The latter means that the optimization problems $\max_{u\in U}\langle a,u\rangle, \quad \max_{v\in V}\langle b,v\rangle$ can be efficiently solved for any a,b. An example of such domains is the unit ball of a norm $\|\cdot\|$ for which the dual norm $\|\cdot\|_*$ is efficiently computable. Let us now demonstrate how an accuracy certificate can be computed in this situation (see (Nemirovski et al., 2010; Harchaoui et al., 2015a) for a more detailed exposition).

A *certificate* is simply a sequence $\lambda^t = (\lambda^t_\tau)_{\tau=1}^t$ of positive weights such that $\sum_{\tau=1}^t \lambda^t_\tau = 1$. Consider the λ^t-average of the iterates z_τ obtained by the algorithm,

$$z^t = [u^t, v^t] = \sum_{\tau=1}^t \lambda^t_\tau z_\tau.$$

A trivial example of certificate corresponds to the constant stepsize, and amounts to simple averaging. However, one might consider other choices of certificate, for which theoretical complexity bounds are preserved – for example, it might be practically reasonable to average only the last portion of the iterates, a strategy called "suffix averaging" (Rakhlin et al., 2012). The point is that any certificate implies a non-trivial (and easily computable) upper bound on the accuracy of the corresponding candidate solution z^t. Indeed, the duality gap of a composite saddle-point problem can be bounded as follows:

$$\begin{aligned}\overline{\phi}(u^t) - \underline{\phi}(v^t) &= \overline{\phi}(u^t) - \phi(u^t,v^t) + \phi(u^t,v^t) - \underline{\phi}(v^t)\\ &= \max_{v\in V}[\phi(u^t,v) - \phi(u^t,v^t)] - \min_{u\in U}[\phi(u,v^t) - \phi(u^t,v^t)]\\ &\le \max_{v\in V}[\phi(u^t,v) - \phi(u^t,v^t)] + \max_{u\in U}[\phi(u^t,v^t) - \phi(u,v^t)].\end{aligned}$$

Now, using concavity of f in v, we have

$$\phi(u^t,v) - \phi(u^t,v^t) = f(u^t,v) - f(u^t,v^t) \le \sum_{\tau=1}^t \lambda^t_\tau \langle F_v(z_\tau), v^t - v\rangle.$$

On the other hand, by convexity of f and Ψ in u,

$$\phi(u^t,v^t) - \phi(u,v^t) = f(u^t,v^t) - f(u,v^t) + \Psi(u^t) - \Psi(u) \le \sum_{\tau=1}^t \lambda^t_\tau \langle F_u(z_\tau) + h(u_\tau), u^t - u\rangle$$

where $h(u_\tau)$ is a subgradient of $\Psi(\cdot)$ at u_τ. Combining the above facts, we get that

$$\overline{\phi}(u^t) - \underline{\phi}(v^t) \le \max_{u\in U}[-F^t_u - h^t] + \max_{v\in V}[-F^t_v] + \sum_{\tau=1}^t \lambda^t_\tau \left[\langle F_u(z_\tau) + h(u_\tau), u^t\rangle + \langle F_v(z_\tau), v^t\rangle\right], \tag{38}$$

where

$$F^t_u = \sum_{\tau=1}^t \lambda^t_\tau F_u(z_\tau), \quad F^t_v = \sum_{\tau=1}^t \lambda^t_\tau F_v(z_\tau), \quad \text{and } \; h^t = \sum_{\tau=1}^t \lambda^t_\tau h(u_\tau).$$

Note that the corresponding averages can often be recomputed in linear time in the dimension of the problem, and then upper bound (38) can be efficiently maintained. For example, this is the case when λ^t corresponds to a fixed sequence $\gamma_1, \gamma_2, ...$,

$$\lambda^t_\tau = \frac{\gamma_\tau}{\sum_{\tau'\le t}\gamma_{\tau'}}, \quad \tau \le t.$$

Note also that any bound on the duality gap implies bounds on the *relative* accuracy for the primal and the dual problem provided that $\underline{\phi}(v^t)$ (and hence the optimal value $\phi(u^*, v^*)$) is strictly positive (we used this fact in our experiments, see Sec. 5). Indeed, let $\varepsilon(t)$ be an upper bound on the duality gap (*e.g.* such as (38)), and hence also on the primal accuracy:

$$\overline{\phi}(u^t) - \phi(u^*, v^*) \le \overline{\phi}(u^t) - \underline{\phi}(v^t) \le \varepsilon(t).$$

Then, since $\phi(u^*, v^*) \ge \underline{\phi}(v^t) > 0$, we arrive at

$$\frac{\overline{\phi}(u^t) - \phi(u^*, v^*)}{\phi(u^*, v^*)} \le \frac{\varepsilon(t)}{\underline{\phi}(v^t)}.$$

A similar bound can be obtained for the relative accuracy of the dual problem.

B. Computation of prox-mappings

It suffices to consider partial proximal setups separately; the case of joint setup in saddle-point problems can be treated using that the joint prox-mapping is separable in u and v, cf. Sec. 2.3. Recall that the possible partial setups $(\|\cdot\|, \omega(\cdot))$ comprise the ℓ_2-setup with $\|\cdot\| = \|\cdot\|_{\mathbb{C},2} = \|\cdot\|_2$ and the (complex) ℓ_1-setup with $\|\cdot\| = \|\cdot\|_{\mathbb{C},1}$; in both cases, $\omega(\cdot)$ is given by (7). Computing $\text{Prox}_{\frac{1}{L}\Psi,u}(g)$, cf. (11), amounts to solving

$$\min_{\xi \in \mathbb{R}^N} \left\{ \xi^{\mathrm{T}}(g - \omega'(u)) + \omega(\xi) : \|\xi\|_{\mathbb{C},q} \le R \right\}, \tag{39}$$

in the constrained case, and

$$\min_{\xi \in \mathbb{R}^N} \left\{ \xi^{\mathrm{T}}(g - \omega'(u)) + \omega(\xi) + \frac{\lambda}{L}\|\xi\|_{\mathbb{C},1}^q \right\}, \tag{40}$$

in the penalized case[3]; in both cases, $q \in \{1, 2\}$. In the constrained case with ℓ_2-setup, the task is reduced to the Euclidean projection onto the ℓ_2-ball if $q = 2$, and onto the ℓ_1-ball if $q = 1$; the latter can be done (exactly) in $\widetilde{O}(N)$ via the algorithm from (Duchi et al., 2008) – for that, one first solves (39) for the complex phases corresponding to the pairs of components of ξ. The constrained case with ℓ_1-setup is reduced to the penalized case by passing to the Langrangian dual problem. Evaluation of the dual function amounts to solving a problem equivalent to (40) with $q = 1$, and (39) can be solved by a simple root-finding procedure if one is able to solve (40). As for (40), below we show how to solve it explicitly when $q = 1$, and reduce it to one-dimensional root search (so that it can be solved in $O(n)$ to numerical tolerance) when $q = 2$. Indeed, (40) can be recast in terms of the complex variable $\zeta = \text{Vec}_n^{\mathrm{H}}\xi$:

$$\min_{\zeta \in \mathbb{C}^{n+1}} \left\{ \langle \zeta, z \rangle + \underline{\omega}(\zeta) + \frac{\lambda}{L}\|\zeta\|_1^q \right\}, \tag{41}$$

where $z = \text{Vec}_n^{\mathrm{H}}(g - \omega'(u))$, and $\underline{\omega}(\zeta) = \omega(\xi)$, cf. (7), whence

$$\underline{\omega}(\zeta) = \frac{C(m, \tilde{q}, \tilde{\gamma})\|\zeta\|_{\tilde{q}}^2}{2}, \tag{42}$$

with $C(m, \tilde{q}, \tilde{\gamma}) = \frac{1}{\tilde{\gamma}}(m+1)^{(\tilde{q}-1)(2-\tilde{q})/\tilde{q}}$. Now, (41) can be minimized first with respect to the complex arguments, and then to the absolute values of the components of ζ. Denoting ζ^* a (unique) optimal solution of (41), the first minimization results in $\zeta_j^* = -\frac{z_j}{|z_j|}|\zeta_j^*|,\ 0 \le j \le n$, and it remains to compute the absolute values $|\zeta_j^*|$.

Case $q = 1$. The first-order optimality condition implies

$$C(m, \tilde{q}, \tilde{\gamma})\|\zeta^*\|_{\tilde{q}}^{2-\tilde{q}}|\zeta_j^*|^{\tilde{q}-1} + \frac{\lambda}{L}\mathbb{1}\{|\zeta_j^*| > 0\} = |z_j|. \tag{43}$$

Denoting $\tilde{p} = \frac{\tilde{q}}{\tilde{q}-1}$, and using the soft-thresholding operator

$$\text{Soft}_M(x) = (|x| - M)_+ \,\text{sign}(x),$$

[3] For the purpose of future reference, we also consider the case of squared $\|\cdot\|_{\mathbb{C},1}$-norm penalty.

we obtain the explicit solution:

$$\zeta_j^* = \frac{1}{C(m,\tilde{q},\tilde{\gamma})}\left(\frac{\theta_j}{\|\theta\|_{\tilde{p}}^{2-\tilde{q}}}\right)^{\tilde{p}/\tilde{q}}, \quad \theta_j = \mathrm{Soft}_{\lambda/L}(z_j).$$

In the case of ℓ_2-setup this reduces to $\zeta_j^* = \mathrm{Soft}_{\lambda/L}(z_j)$.

Case $q = 2$. Instead of (43), we arrive at

$$C(m,\tilde{q},\tilde{\gamma})\|\zeta^*\|_{\tilde{q}}^{2-\tilde{q}}|\zeta_j^*|^{\tilde{q}-1} + \frac{2\lambda\|\zeta^*\|_1}{L}\mathbb{1}\{|\zeta_j^*| > 0\} = |z_j|, \tag{44}$$

which we cannot solve explicitly. However, note that a counterpart of (44), in which $\|\zeta^*\|_1$ is replaced with parameter $t \ge 0$, can be solved explicitly similarly to (43). Let $\zeta^*(t)$ denote the corresponding solution for a fixed t, which can be obtained in $O(n)$ time. Clearly, $\|\zeta^*(t)\|_1$ is a non-decreasing function on $\mathbb{R}_+$. Hence, (44) can be solved, up to numerical tolerance, by any one-dimensional root search procedure, in $O(1)$ evaluations of $\zeta^*(t)$.

C. Technical proofs

Proof of Lemma 4.1. Note that $\mathcal{A}$ can be expressed as follows, cf. (18):

$$\mathcal{A} = \sqrt{2n+1}\cdot F_n P_n F_{2n}^{\mathrm{H}} D_y F_{2n} P_n^{\mathrm{H}} F_n^{\mathrm{H}}. \tag{45}$$

By Young's inequality, for any $\psi \in \mathbb{C}^{n+1}$ we get

$$\begin{aligned}\frac{1}{2n+1}\|\mathcal{A}\psi\|_2^2 &\le \left\|D_y F_{2n} P_n^{\mathrm{H}} F_n^{\mathrm{H}}\psi\right\|_2^2 \\ &\le \left\|F_{2n}[y]_{-n}^n\right\|_\infty^2 \left\|F_{2n} P_n^{\mathrm{H}} F_n^{\mathrm{H}}\psi\right\|_2^2 \\ &\le \left\|F_{2n}[y]_{-n}^n\right\|_\infty^2 \|\psi\|_2^2,\end{aligned}$$

where we used that P_n is non-expansive. □

Proof of Proposition 4.2. Consider the uniform grid on the unit circle

$$U_n = \left\{\exp\left(\frac{2\pi ij}{n+1}\right)\right\}_{j=0}^n,$$

and the twice finer grid

$$U_N = \left\{\exp\left(\frac{2\pi ij}{N+1}\right)\right\}_{j=0}^N, \quad N = 2n+1.$$

Note that U_N is the union of U_n and the shifted grid

$$\tilde{U}_n = \left\{u\, e^{i\theta},\ u \in U_n\right\}, \quad \theta = \frac{2\pi}{N+1};$$

note that $\tilde{U}_n$ and U_n do not overlap. One can check that for any $n \in \mathbb{Z}_+$ and $x \in \mathbb{C}_n(\mathbb{Z})$, the components of $F_n[x]_0^n$ form the set

$$\left\{\frac{x(\nu)}{\sqrt{n+1}}\right\}_{\nu \in U_n},$$

where $x(\cdot)$ is the Taylor series corresponding to x:

$$x(\nu) := \sum_{\tau\in\mathbb{Z}} x_\tau \nu^\tau.$$

Now, let x be as in the premise of the theorem, and let $x^{(n)} \in \mathbb{C}_n(\mathbb{Z})$ be such that $x^{(n)}_\tau = x_\tau$ if $0 \le \tau \le n$ and $x^{(n)}_\tau = 0$ otherwise. Similarly, let us introduce $x^{(N)}$ as x restricted on $\mathbb{C}_N(\mathbb{Z})$. Then one can check that for any $\nu \in U_N$,

$$x^{(N)}(\nu) = \begin{cases} 2x^{(n)}(\nu), & \nu \in U_n, \\ 0, & \nu \in \tilde{U}_n. \end{cases} \tag{46}$$

In particular, this implies that

$$\|F_N[x]_0^N\|_\infty = \sqrt{2}\|F_n[x]_0^n\|_\infty. \tag{47}$$

Now, for any $\varphi \in \mathbb{C}_n(\mathbb{Z})$, let $\phi \in \mathbb{C}_N(\mathbb{Z})$ be its $n+1$-periodic extension, defined by

$$[\phi]_0^N = [[\varphi]_0^n; [\varphi]_0^n].$$

One can directly check that for x as in the premise of the theorem, the circular convolution of $[\phi]_0^N$ and $[x]_0^N$ is simply a one-fold repetition of $2[\varphi * x]_0^n$. Hence, using the Fourier diagonalization property together with (47) applied for $[\varphi * x]_0^n$ instead of x_0^n, we obtain

$$\sqrt{N+1}\,\|F_N[x] \odot F_N[\phi]\|_\infty = 2\sqrt{2}\,\|F_n[x * \varphi]\|_\infty \tag{48}$$

where $a \odot b$ is the elementwise product of $a, b \in \mathbb{C}^{n+1}$.

Finally, note that since $\sigma = 0$, and, as such, $x = y$ a.s., for any $\psi \in \mathbb{C}^{n+1}$ one has:

$$\mathcal{A}\psi = F_n[x * \varphi], \quad \text{where } \varphi = F_n^{\mathrm{H}}[\psi] \in \mathbb{C}_n(\mathbb{Z}).$$

Hence, using (48) with such φ, we arrive at

$$\begin{aligned} \|\mathcal{A}\psi\|_\infty &= \|F_n[x * \varphi]\|_\infty \\ &= \frac{\sqrt{n+1}}{2}\|F_N[x] \odot F_N[\phi]\|_\infty && \text{[by (48)]} \\ &= \sqrt{n+1}\|F_n[x] \odot \psi\|_\infty. && \text{[by (46)]} \end{aligned}$$

The claim now follows by maximizing the right-hand side in $\psi \in \mathbb{C}^{n+1} : \|\psi\|_1 \le 1$. □

C.1. Proof of Theorem 4.3

The proof is reduced to the following observation: in order to satisfy (24), it suffices for $\tilde{\varphi} \in \mathbb{C}_n(\mathbb{Z})$ to satisfy

$$\|F_n\tilde{\varphi}\|_1 = O\left(\frac{r}{\sqrt{n+1}}\right), \quad \|F_n[y - y * \tilde{\varphi}]\|_\infty = \tilde{O}\left(\sigma r\right), \text{ where } r = 2\rho^2.$$

This is a rather straightforward remark to the proof of Proposition 4 in (Harchaoui et al., 2015b). We give here the proof for convenience of the reader, and also consider the case of the penalized estimator.

Preliminaries. Let Δ be the unit lag operator such that $[\Delta x]_t = x_{t-1}$ for $x \in \mathbb{C}(\mathbb{Z})$. Note that for any filter $\varphi \in \mathbb{C}_n(\mathbb{Z})$, one can write $\varphi * y = \varphi(\Delta)y$ where $\varphi(\Delta)$ is the Taylor polynomial corresponding to φ:

$$\varphi(\Delta) := \sum_{\tau\in\mathbb{Z}} \varphi_\tau \Delta^\tau = \sum_{0\le\tau\le n} \varphi_\tau \Delta^\tau.$$

Besides, let us introduce the random variable

$$\Theta_n(\zeta) := \max_{0\le\tau\le n} \|\Delta^\tau F_n[\zeta]\|_\infty.$$

Note that $F_n[\zeta]$ is distributed same as $[\zeta]_0^n$ by the unitary invariance of the law $\mathbb{C}\mathcal{N}(0, I_n)$. Using this fact, it is straightforward to obtain that with probability at least $1-\delta$,

$$\Theta_n(\zeta) \le \overline{\Theta}_n := 4\sqrt{\log\left(\frac{n+1}{\delta}\right)}, \tag{49}$$

see (Harchaoui et al., 2015b).

Constrained uniform-fit estimator. Let $\widehat{\varphi}$ be an optimal solution to (Con-UF) with $\overline{r} = r$. We begin with the following decomposition (recall that $\varphi * y = \varphi(\Delta)y$):

$$\begin{aligned}
|[x - \widehat{\varphi}(\Delta)y]_n| &\le \sigma|[\widehat{\varphi}(\Delta)\zeta]_n| + |[x - \widehat{\varphi}(\Delta)x]_n| \\
&\le \sigma\|F_n[\widehat{\varphi}]\|_1\|F_n[\zeta]\|_\infty + |[x - \widehat{\varphi}(\Delta)x]_n| \\
&\le \frac{\sigma r\Theta_n(\zeta)}{\sqrt{n+1}} + |[x - \widehat{\varphi}(\Delta)x]_n|.
\end{aligned} \tag{50}$$

Here, to obtain the second line we used Young's inequality, and for the last line we used feasibility of $\widehat{\varphi}$ in (Con-UF). Now let us bound $|[x - \widehat{\varphi}(\Delta)x]_n|$:

$$\begin{aligned}
|[x - \widehat{\varphi}(\Delta)x]_n| &\le |[(1 - \widehat{\varphi}(\Delta))(1 - \varphi^o(\Delta))x]_n| \; + \; |[\varphi^o(\Delta)(1 - \widehat{\varphi}(\Delta))x]_n| \\
&\le (1 + \|\widehat{\varphi}\|_1)\,\|[(1 - \varphi^o(\Delta))x]_0^n\|_\infty \; + \; \|F_n[\varphi^o]\|_1\|F_n[(1 - \widehat{\varphi}(\Delta))x]\|_\infty.
\end{aligned}$$

Discrepancy of the oracle φ^o in the time domain can be bounded using (36):

$$\|[(1 - \varphi^o(\Delta))x]_0^n\|_\infty \le \frac{4r\sigma}{\sqrt{n+1}}. \tag{51}$$

Indeed, for any $\tau \in \mathbb{Z}$, $[(1 - \varphi^o(\Delta))x]_\tau = [x - \varphi^o(\Delta)y]_\tau + \sigma[\varphi^o(\Delta)\zeta]_\tau$. On the other hand, using that φ^o is non-random,

$$\mathbf{E}|[\varphi^o(\Delta)\zeta]_\tau|^2 = \|\varphi^o\|_2^2 = \|F_n[\varphi^o]\|_2^2 \le \|F_n[\varphi^o]\|_1^2 = \frac{r^2}{n+1}.$$

Now, using that due to (37) oracle φ^o is feasible in (Con-UF), we can bound the Fourier-domain discrepancy of $\widehat{\varphi}$:

$$\begin{aligned}
\|F_n[(1 - \widehat{\varphi}(\Delta))x]\|_\infty &\le \left\|F_n[(1 - \widehat{\varphi}(\Delta))y]\right\|_\infty + \sigma\|F_n[(1 - \widehat{\varphi}(\Delta))\zeta]\|_\infty \\
&\le \left\|F_n[(1 - \widehat{\varphi}(\Delta))y]\right\|_\infty + \sigma(1 + \|\widehat{\varphi}\|_1)\Theta_n(\zeta) \\
&\le \left\|F_n[(1 - \varphi^o(\Delta))y]\right\|_\infty + \sigma(1 + \|\widehat{\varphi}\|_1)\Theta_n(\zeta)
\end{aligned}$$

$$\leq \|F_n[(1-\varphi^o(\Delta))x]\|_\infty + \sigma(2+\|\varphi^o\|_1+\|\widehat{\varphi}\|_1)\Theta_n(\zeta). \tag{52}$$

Meanwhile, using (51), we can bound the Fourier-domain discrepancy of φ^o:

$$\begin{aligned}\|F_n[(1-\varphi^o(\Delta))x]\|_\infty &\leq \|F_n[(1-\varphi^o(\Delta))x]\|_2 \\ &= \|[(1-\varphi^o(\Delta))x]_0^n\|_2 \leq 4\sigma r.\end{aligned} \tag{53}$$

Collecting the above, we obtain

$$|[x-\widehat{\varphi}(\Delta)x]_n| \leq (1+\|\widehat{\varphi}\|_1)\frac{4r\sigma}{\sqrt{n+1}} + \sigma\|F_n[\varphi^o]\|_1 \left\{4r + (2+\|\varphi^o\|_1+\|\widehat{\varphi}\|_1)\Theta_n(\zeta)\right\}.$$

Note that $\|F_n[\varphi^o]\|_1$ is bounded by (37). It remains to bound $\|\varphi^o\|_1$ and $\|\widehat{\varphi}\|_1$:

$$\|\varphi^o\|_1 \leq \sqrt{n+1}\|\varphi^o\|_2 \leq \sqrt{n+1}\|F_n[\varphi^o]\|_1 \leq r, \tag{54}$$

and similarly $\|\widehat{\varphi}\|_1 \leq r$. Hence, we have

$$|[x-\widehat{\varphi}(\Delta)x]_n| \leq \frac{\sigma r}{\sqrt{n+1}}\left[4(1+2r)+2(1+r)\Theta_n(\zeta)\right],$$

and, using (50) and (49), we arrive that with probability $\geq 1-\delta$,

$$|x_n - [\widehat{\varphi}(\Delta)y]_n| \leq \frac{C\sigma r^2\sqrt{1+\log\left(\frac{n+1}{\delta}\right)}}{\sqrt{n+1}}. \tag{55}$$

It is now straightforward to see why $\tilde{\varphi}$, an $O(\sigma r)$-accurate solution to (Con-UF), also satisfies (55): the first change in the above argument when replacing $\widehat{\varphi}$ with $\tilde{\varphi}$ is the additional term $O(\sigma r)$ in (52). Since all the remaining terms in the right-hand side of (52) were also bounded from above by $O(\sigma r)$, (55) is preserved for $\tilde{\varphi}$ up to a constant factor. □

Penalized uniform-fit estimator. Let now $\widehat{\varphi}$ be an optimal solution to (Pen-UF). The proof goes along the same lines as in the previous case; however, we must take into account a different condition for oracle feasibility. Proceeding as in (50) and using (51), we get

$$\begin{aligned}&|[x-\widehat{\varphi}(\Delta)y]_n| \\ &\leq \sigma\|F_n[\widehat{\varphi}]\|_1\|F_n[\zeta]\|_\infty + |[(1-\widehat{\varphi}(\Delta))x]_n| \\ &\leq \sigma\|F_n[\widehat{\varphi}]\|_1\|F_n[\zeta]\|_\infty + \|F_n[\varphi^o]\|_1\|F_n[(1-\widehat{\varphi}(\Delta))x]\|_\infty + (1+\|\widehat{\varphi}\|_1)\|[(1-\varphi^o(\Delta))x]_0^n\|_\infty \\ &\leq \sigma\|F_n[\widehat{\varphi}]\|_1\Theta_n(\zeta) + \frac{r}{\sqrt{n+1}}\|F_n[(1-\widehat{\varphi}(\Delta))x]\|_\infty + \frac{4r\sigma}{\sqrt{n+1}}(1+\|\widehat{\varphi}\|_1).\end{aligned} \tag{56}$$

Let us condition on the event $\Theta_n(\zeta) \leq \overline{\Theta}_n$ the probability of which is $\geq 1-\delta$. Feasibility of $\widehat{\varphi}$ in (Pen-UF) yields

$$\begin{aligned}\|F_n[(1-\widehat{\varphi}(\Delta))y]\|_\infty + \lambda\|F_n[\widehat{\varphi}]\|_1 &\leq \|F_n[(1-\varphi^o(\Delta))y]\|_\infty + \lambda\|F_n[\varphi^o]\|_1 \\ &\leq 4\sigma r + (1+r)\sigma\Theta_n(\zeta) + \frac{\lambda r}{\sqrt{n+1}} \\ &\leq \left(4+2\Theta_n(\zeta)+\frac{\lambda}{\sigma\sqrt{n+1}}\right)\sigma r \\ &\leq \frac{2\lambda r}{\sqrt{n+1}}.\end{aligned} \tag{57}$$

Here first we used (53), (54), and the last line of (52), then that $r \geq 1$, and, finally, used the choice of λ from the premise of the theorem. Now from (57) we obtain

$$\|F_n[\widehat{\varphi}]\|_1 \leq \frac{2r}{\sqrt{n+1}} \tag{58}$$

and

$$1+\|\widehat{\varphi}\|_1 \le 1+\sqrt{n+1}\|F_n[\widehat{\varphi}]\|_1 \le 1+2r \le 3r. \tag{59}$$

Further, using (57) and (59), we get

$$\begin{aligned}\|F_n[(1-\widehat{\varphi}(\Delta))x]\|_\infty &\le \|F_n[(1-\widehat{\varphi}(\Delta))y]\|_\infty + \sigma(1+\|\widehat{\varphi}\|_1)\Theta_n(\zeta)\\ &\le \left(\frac{2\lambda}{\sigma\sqrt{n+1}} + 3\Theta_n(\zeta)\right)\sigma r \end{aligned} \tag{60}$$

Substituting (58)–(60) into (56), we arrive at

$$\begin{aligned}|[x-\widehat{\varphi}(\Delta)y]_0| &\le \left(\frac{2\lambda}{\sigma\sqrt{n+1}} + 5\Theta_n(\zeta)+8\right)\frac{\sigma r^2}{\sqrt{n+1}}\\ &\le \frac{5\lambda r^2}{n+1}\\ &= \frac{80r^2\sqrt{1+\log\left(\frac{n+1}{\delta}\right)}}{\sqrt{n+1}}.\end{aligned}$$

Similarly to the case of the constrained estimator, it is straightforward to see that the last bound is preserved (up to a constant factor) for an ε-accurate solution $\tilde{\varphi}$ to (Pen-UF) with $\varepsilon = O(\sigma r)$. □

C.2. Proof of Theorem 4.4

Constrained least-squares estimator. Let us first summarize the original proof of (25) for the case of an exact optimal solution $\widehat{\varphi}$ of (Con-LS), see Theorem 2.2 in (Ostrovsky et al., 2016) and its full version (Ostrovsky et al.). Introducing the scaled Hermitian dot product for $\varphi, \psi \in \mathbb{C}_n(\mathbb{Z})$,

$$\langle \varphi, \psi \rangle_n = \frac{1}{n+1}\sum_{\tau=0}^{n} \overline{\varphi}_\tau \psi_\tau,$$

the squared ℓ_2-loss can be decomposed as follows:

$$\begin{aligned}\|x-\widehat{\varphi}*y\|_{n,2}^2 &= \|y-\widehat{\varphi}*y\|_{n,2}^2 - \sigma^2\|\zeta\|_{n,2}^2 - 2\sigma\langle\zeta, x-\widehat{\varphi}*y\rangle_n\\ &\le \|y-\varphi^o*y\|_{n,2}^2 - \sigma^2\|\zeta\|_{n,2}^2 - 2\sigma\langle\zeta, x-\widehat{\varphi}*y\rangle_n\\ &= \|x-\varphi^o*y\|_{n,2}^2 + 2\sigma\langle\zeta, x-\varphi^o*y\rangle_n - 2\sigma\langle\zeta, x-\widehat{\varphi}*y\rangle_n,\end{aligned} \tag{61}$$

where the inequality is due to feasibility of φ^o in (Con-LS). Now, it turns out that the dominating term in the right-hand side is the first one (corresponding to the squared oracle loss): we know that due to (36), with probability $\ge 1-\delta$ one has

$$\|x-\varphi^o*y\|_{n,2}^2 \le \frac{9\sigma^2 r^2\log\left(\frac{n+1}{\delta}\right)}{n+1}. \tag{62}$$

On the other hand, one can bound the next term in the right-hand side of (61) as

$$\begin{aligned}\sigma\langle\zeta, x-\varphi^o*y\rangle_n &\le \frac{\sigma\sqrt{2\log\left(\frac{3}{\delta}\right)}}{\sqrt{n+1}}\|x-\varphi^o*y\|_{n,2} + \frac{12\sigma^2 r(1+\log\left(\frac{6}{\delta}\right))}{n+1}\\ &\le \frac{6\sigma^2 r\log\left(\frac{3(n+1)}{\delta}\right)}{n+1} + \frac{12\sigma^2 r(1+\log\left(\frac{6}{\delta}\right))}{n+1}\\ &\le \frac{30\sigma^2 r\log\left(\frac{6(n+1)}{\delta}\right)}{n+1}.\end{aligned} \tag{63}$$

Here, for the first inequality we refer the reader to the original proof in (Ostrovsky et al.), eq. (44-45), where one should set $\kappa_{m,n}=1$ and keep in mind the absence of scaling factor $\frac{1}{n+1}$ in the definitions of $\langle\phi,\psi\rangle_n$ and $\|\cdot\|_{n,2}$. The next inequalities then follow by simple algebra using (62).

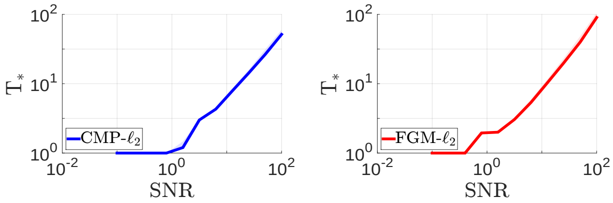

Figure 4: Iteration at which the accuracy ε_* is attained for (Con-UF), left, and (Con-LS), right, in *Random*-4.

Finally, the last term in the right-hand side of (61) can be bounded as follows with probability $\geq 1-\delta$:

$$2\sigma|\langle\zeta, x-\widehat{\varphi}*y\rangle_n| \leq \frac{2\sqrt{2}\sigma\left(\sqrt{s}+\sqrt{\log\left(\frac{2}{\delta}\right)}\right)}{\sqrt{n+1}}\|x-\widehat{\varphi}*y\|_{n,2} + \frac{8\sqrt{2}\sigma^2 r\left(2+\log\left(\frac{8(n+1)}{\delta}\right)\right)}{n+1}. \tag{64}$$

see eq. (33-40) in (Ostrovsky et al.) where one must set $\varkappa = 0$ in our setting since $x \in \mathcal{S}$. Moreover, in the proof of (64) the optimality of $\widehat{\varphi}$ was not used; instead, the argument in (Ostrovsky et al.) relied only on the following facts:

(i) $x \in \mathcal{S}$ where $\mathcal{S}$ is a shift-invariant subspace of $\mathbb{C}(\mathbb{Z})$ with $\dim(\mathcal{S}) = s$;

(ii) one has a bound on the Fourier-domain ℓ_1-norm of $\widehat{\varphi}$: $\|F_n[\widehat{\varphi}]\|_1 \leq \frac{r}{\sqrt{n+1}}$.

Finally, collecting (61)-(64) and solving the resulting quadratic inequality, one bounds the scaled ℓ_2-loss of $\widehat{\varphi}$:

$$\|x-\widehat{\varphi}*y\|_{n,2} \leq \frac{C\sigma}{\sqrt{n+1}}\left(\sqrt{s}+r\sqrt{\log\left(\frac{n+1}{\delta}\right)}\right). \tag{65}$$

(We used that $r \geq 1$.) Moreover, it is now evident that an ε-accurate solution $\widehat{\varphi}$ to (Con-LS) with $\varepsilon = O(\sigma^2 r^2)$ still satisfies (65). Indeed, the error decomposition (61) must now be replaced with

$$\|x-\tilde{\varphi}*y\|_{n,2}^2 \leq \|x-\varphi^o*y\|_{n,2}^2 + 2\sigma\langle\zeta, x-\varphi^o*y\rangle_n - 2\sigma\langle\zeta, x-\tilde{\varphi}*y\rangle_n + \frac{\varepsilon}{n+1}. \tag{66}$$

Then, (62) and (63) do not depend on $\tilde{\varphi}$, and hence are preserved. The term $\frac{\varepsilon}{n+1}$ enters additively, and allows for the same upper bound as (62). Finally, (64) is preserved when replacing $\widehat{\varphi}$ with $\tilde{\varphi}$ since (i) and (ii) remain true. □

Penalized least-squares estimator. Let now $\tilde{\varphi}$ be an ε-accurate solutions to (Pen-LS), let $\lambda_n = \frac{\lambda}{\sqrt{n+1}}$, and let $\tilde{r} = \sqrt{n+1}\|F_n[\tilde{\varphi}]\|_1$. Similarly to (66), one has

$$\|x-\tilde{\varphi}*y\|_{n,2}^2 \leq \|x-\varphi^o*y\|_{n,2}^2 + 2\sigma\langle\zeta, x-\varphi^o*y\rangle_n - 2\sigma\langle\zeta, x-\tilde{\varphi}*y\rangle_n + \frac{\lambda_n(r-\tilde{r})}{n+1} + \frac{\varepsilon}{n+1}. \tag{67}$$

Note that (62) and (63) are still valid. Moreover, (64) is preserved for $\tilde{\varphi}$ if r is replaced with $\tilde{r}$, cf. (i) and (ii):

$$2\sigma|\langle\zeta, x-\tilde{\varphi}*y\rangle_n| \leq \frac{2\sqrt{2}\sigma\left(\sqrt{s}+\sqrt{\log\left(\frac{2}{\delta}\right)}\right)}{\sqrt{n+1}}\|x-\tilde{\varphi}*y\|_{n,2} + \frac{8\sqrt{2}\sigma^2 \tilde{r}\left(2+\log\left(\frac{8(n+1)}{\delta}\right)\right)}{n+1}. \tag{68}$$

Hence, if λ is chosen as in the premise of the theorem, the second term in the right-hand side is dominated by $\frac{\lambda_n\tilde{r}}{n+1}$. Combining (62), (63), and (68) with the fact that $\varepsilon = O(\sigma^2 r^2)$, plugging in the value of λ from the premise of the theorem, and solving the resulting quadratic inequality, we conclude that (65) is preserved for $\tilde{\varphi}$. □

D. Additional experiments

Statistical Complexity Bound. In this experiment (see Fig. 4), we illustrate the affine dependency of the statistical complexity T_* from SNR predicted by our theory, see (26) and (27); note that although the signal in *Random* is not sparse on the DFT grid, its DFT is likely to have only a few large spikes which would suffice for (27). For various SNR values, we generate a signal in scenario *Random*-4, and define the first iteration at which $\varepsilon(T)$ crosses level σr for (Con-UF) solved with Algorithm 2, and $\sigma^2 r^2$ for (Con-LS) with Algorithm 1. We see that the log-log curves plateau for low SNR and have unit tangent for high SNR, confirming our predictions.

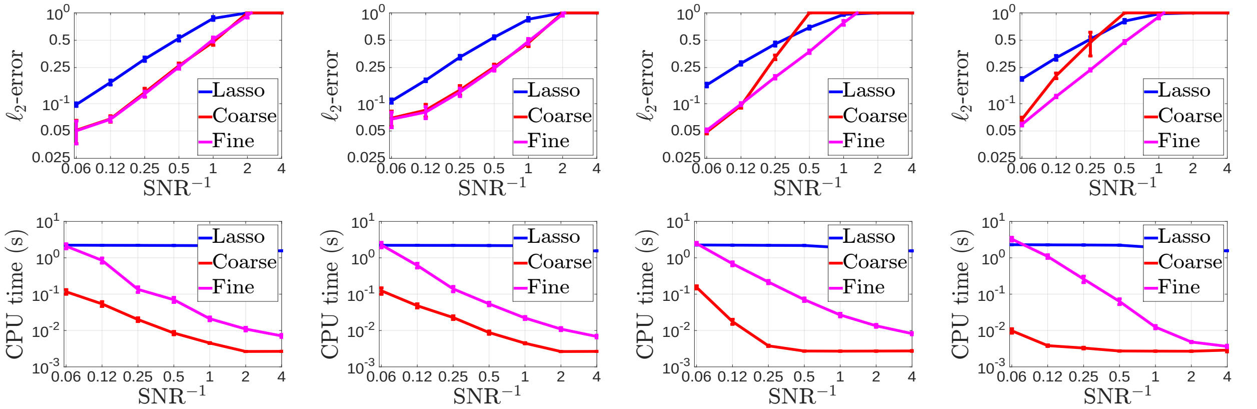

Figure 5: ℓ_2-loss and CPU time spent to compute estimators φ^{coarse}, φ^{fine}, and Lasso.

Statistical Performance with Early Stopping. In this experiment, we present additional scenario *Modulated*-s-m, in which the signal is a sum of sinusoids with polynomial modulation: $x_t = \sum_{k=1}^{s} p_k(t) e^{i\omega_k t}$, where $p_k(\cdot)$ are i.i.d. polynomials of degree r with i.i.d. coefficients sampled from $\mathbb{C}\mathcal{N}(0,1)$; note that in this case $\dim(\mathcal{S}) = 2s(m+1)$. Our goal is to study how the early stopping of an algorithm upon reaching accuracy ε_* (using an accuracy certificate) affects the statistical performance of the resulting estimator. For that, we generate signals in scenarios *Random*-4, *Coherent*-2, *Modulated*-4-2 (quadratic modulation), and *Modulated*-4-4 (quartic modulation), with different SNR, and compare three estimators: approximate solution φ^{coarse} to (Con-LS) with guaranteed accuracy $\varepsilon_* = \sigma^2 r^2$, near-optimal solution φ^{fine} with guaranteed accuracy $0.01\varepsilon_*$, and the Lasso estimator, with the standard choice of parameters as described in (Bhaskar et al., 2013), which we compute by running 3000 iterations of the FISTA algorithm (Beck & Teboulle, 2009); note that the optimization problem in the latter case is unconstrained, and we do not have an accuracy certificate. We plot the scaled ℓ_2-loss of an estimator and the CPU time spent to compute it (we used MacBook Pro 2013 with 2.4 GHz Intel Core i5 CPU and 8GB of RAM). The results are shown in Fig. 5. We observe that φ^{coarse} has almost the same performance as φ^{fine} while being computed 1-2 orders of magnitude faster on average; both significantly outperform Lasso in all scenarios.