Proof of Theorem 1

Regarding $\mathcal{L}$, we assume it is continuously differentiable in the N-dimensional real space, i.e. $\mathcal{L} \in C^1(R^N)$.

Theorem A1: Suppose that there exists $L > 0$ so that

$$||\nabla\mathcal{L}(\theta_1; x) - \nabla\mathcal{L}(\theta_2; x)||_2 \leq L||\theta_1 - \theta_1||_2, \forall\theta_1, \theta_2 \in R^N. \tag{1}$$

Then, for all θ_1 and $\theta_2 \in R^N$, it holds that [1]:

$$\mathcal{L}(\theta_2; x) \leq \mathcal{L}(\theta_1; x) + \nabla\mathcal{L}(\theta_1)^T(\theta_2 - \theta_1) + \frac{L}{2}||\theta_2 - \theta_1||_2^2. \tag{2}$$

Generally, $L = ||\frac{\partial\mathcal{L}}{\partial\theta}||_{2,max}$. Due to the use of L1 regularization, $\frac{\partial x}{\partial\theta}$ will be very sparse and v_t is bounded by $\frac{G^2}{1-\beta}$, where $G = ||\frac{\partial\mathcal{L}}{\partial x}||_{2,max}$.

Lemma A1: If $||\frac{\partial x}{\partial\theta}||_2 \ll ||\frac{\partial\mathcal{L}}{\partial x^T}||_2$, then $||\frac{\partial\mathcal{L}}{\partial\theta}||_{2,max} \leq ||\frac{\partial\mathcal{L}}{\partial x^T}||_2 \leq ||\sqrt{v_t}|| \leq ||\frac{\sqrt{v_t}}{\eta}||_2$.

According to **Lemma A1**, L can be set as $||\frac{\sqrt{v_t}}{\eta}||_2$.

Let $\theta \in$ dom g. Note from **Theorem A1** that we have for all $k \geq 0$

$$\begin{aligned} F(\theta_{k+1}; x) \leq & \mathcal{L}(\theta_k; x) + \nabla\mathcal{L}(\theta_k; x)^T(\theta_{k+1} - \theta_k) \\ & + \frac{1}{2\eta}||\theta_{k+1} - \theta_k||_2^2 + ||\theta_{k+1}||_1, \end{aligned} \tag{3}$$

Theorem A2: Let F be proper, closed, and strongly convex, for any $\theta \in R^n$, it holds that

$$\frac{\sigma}{2}||\theta - \hat{\theta}||_2^2 \leq F(\theta) - F(\hat{\theta}), \tag{4}$$

where $\sigma > 0$ and is called a strong convexity modulus. Generally, under the condition of **Theorem A1**, σ can be set $\frac{\sqrt{v_t}}{\eta}$. For all $k \geq 0$

$$\begin{aligned} F(\theta_{k+1}; x) \leq & \mathcal{L}(\theta_k; x) + \nabla\mathcal{L}(\theta_k; x)^T(\theta - \theta_k) \\ & + \frac{1}{2\eta}||\theta - \theta_k||_2^2 + ||\theta||_1 - ||\frac{\sqrt{v_t}}{\eta}||_2||\theta_{k+1} - \theta||_2^2. \end{aligned} \tag{5}$$

Setting $x = x^k$, we get

$$F(\theta_{k+1}; x) \leq \mathcal{L}(\theta_k; x) + ||\theta||_1 - ||\frac{\sqrt{v_t}}{\eta}||_2||\theta_{k+1} - \theta_k||_2^2 \leq F(\theta_k; x), \tag{6}$$

showing that $F(\theta_k; x)$ is nonincreasing. Now, pick any $\hat{\theta} \in$ Arg min F and let $\theta = \hat{\theta}$, then

$$\begin{aligned} F(\theta_{k+1}; x) \leq & \mathcal{L}(\theta_k; x) + \nabla\mathcal{L}(\theta_k; x)^T(\hat{\theta} - \theta_k) \\ & + \frac{1}{2\eta}||\hat{\theta} - \theta_k||_2^2 + ||\hat{\theta}||_1 - ||\frac{\sqrt{v_t}}{\eta}||_2||\theta_{k+1} - \hat{\theta}||_2^2 \\ \leq & \mathcal{L}(\hat{\theta}; x) + ||\hat{\theta}||_1 + \frac{1}{2\eta}||\hat{\theta} - \theta_k||_2^2 - ||\frac{\sqrt{v_t}}{\eta}||_2||\theta_{k+1} - \hat{\theta}||_2^2. \end{aligned} \tag{7}$$

Thus, we obtain:

$$
\begin{aligned}
&(k+1)[F(\theta_{k+1};x)-F(\hat{\theta};x)] \leq \sum_{i=0}^{k}(F(\theta_{i+1};x)-F(\hat{\theta})) \\
&\leq \frac{1}{2\eta}\sum_{i=0}^{k}||\hat{\theta}-\theta_i||_2^2-||\theta_{i+1}-\hat{\theta}||_2^2 \leq ||\frac{\sqrt{v_t}}{\eta}||_2||\hat{\theta}-\theta_0||_2^2 \leq \frac{G}{\sqrt{1-\beta}\eta}||\hat{\theta}-\theta_0||_2^2.
\end{aligned}
\tag{8}
$$

Hence,

$$
F(\theta_k;x)-F(\hat{\theta}) \leq \frac{G}{2k\eta\sqrt{1-\beta}}||\theta_0-\hat{\theta}||_2^2. \tag{9}
$$

The Algorithm's Robustness

Consider an ANN represented as $f(u|\theta)$, where u denotes the input, θ the network parameters, and $x = f(u|\theta)$. When noise ϵ perturbs the observed variables, the actual input becomes $\hat{u} = u + \epsilon$, where ϵ is assumed to follow a normal distribution $\mathcal{N}(0, 0.01)$, following the same setting as in [2]. The resulting output can then be approximated via a first-order expansion: $\hat{x} = f(\hat{u}|\theta) = f(u+\epsilon|\theta) \approx f(u|\theta) + \frac{\partial x}{\partial u}^T \epsilon$.

Taking the L_2-norm of the perturbation term, we obtain: $\left\|\frac{\partial x}{\partial u}^T \epsilon\right\|_2 \leq \left\|\frac{\partial x}{\partial u}\right\|_2 \cdot \|\epsilon\|_2 = \|\theta\|_2 \cdot \|\epsilon\|_2$. Since the proposed method employs L_1-regularization, the parameter matrix θ is highly sparse, leading to a small $\|\theta\|_2$. Additionally, the expected magnitude of the noise term satisfies $\mathbb{E}[\|\epsilon\|_2] = 0.01 \cdot \sqrt{n}$, where n is the number of state variables. This ensures that even for high-dimensional systems, the cumulative effect of noise remains bounded, further reinforcing the algorithm's robustness.

References

[1] S. Boyd and L. Vandenberghe, *Convex optimization.* Cambridge university press, 2004.

[2] C. Pei, Y. Xiao, W. Liang, and *et al.*, "Canonical variate analysis for detecting false data injection attacks in alternating current state estimation," *IEEE Trans. Network Sci. Eng.*, vol. 11, no. 4, pp. 3332–3345, 2024.