Smoothing Property of Load Variation Promotes Finding Global Solutions of Time-Varying Optimal Power Flow

Julie Mulvaney-Kemp, Salar Fattahi, and Javad Lavaei

Abstract

This paper analyzes solution trajectories for optimal power flow (OPF) with time-varying load. Despite its nonlinearity, time-varying OPF is commonly solved every 5-15 minutes using local-search algorithms. Failing to obtain the globally optimal solution of power optimization problems jeopardizes the grid's reliability and causes financial and environmental issues. The objective of this paper is to address this problem by understanding the optimality behavior of OPF solution trajectories. An empirical study on California data shows that, with enough variation in the data, local search methods can solve OPF to global optimality, even if the problem has many local minima. To explain this phenomenon, we introduce a backward mapping that relates the time-varying OPF's global solution at a given time to a set of desirable initial points. We show that this mapping could act as a stochastic gradient ascent algorithm on an implicitly convexified formulation of OPF, justifying the escape of poor solutions over time. This work is the first to mathematically explain how temporal data variation affects the complexity of solving power operational problems.

Index Terms

Optimal power flow, online optimization, global minima, local search

The authors are with the Department of Industrial Engineering and Operations Research, University of California, Berkeley, Berkeley, CA 94709 USA. e-mail: julie_mulvaney-kemp@berkeley.edu; fattahi@berkeley.edu; lavaei@berkeley.edu

A shorter version of this paper has been accepted to the 2020 IEEE Power & Energy Society General Meeting [1]. The new additions to this version include a case study with 16 feasible trajectories (compared with 4 in the previous version), numerical analysis highlighting the impact of load variation, an updated model for approximating time-varying OPF, and theoretical analysis of the problem dynamics over time.

I. Introduction

Optimal power flow (OPF) is a large-scale optimization problem at the core of the daily operation of power systems world-wide. OPF aims to find a cost-minimizing operating point for a power system, subject to various operational and security constraints [2]. The OPF problem is challenging because of its nonconvexity and the frequency at which it is solved [3]. Because demand across the system is constantly in flux, the OPF problem is solved every few minutes to match the system's power generation with its latest demand profile. Nonconvex constraints in the AC model of OPF are the main impediment to solving the problem efficiently and optimally. Physical laws govern these constraints, indicating nonconvexity is inherent to the problem. In power systems [4], [5] and in machine learning [6], such nonconvexity is known to give rise to poor local solutions. To realize the vision of sustainable and resilient power grids, there is a pressing need to address the nonconvexity and timescale of both existing and emerging optimization problems for the control and operation of the grid. Since these problems are all built upon the power flow equations, we focus on OPF in this paper.

With the goal of addressing the underlying nonconvexity of the problem, a recent line of research has focused on approximating the problem as a single or sequence of convex optimization problems. These works include quadratic convex [7], second-order conic programming [8], and semidefinite programming [9]–[11] relaxations. Despite desirable theoretical guarantees, the convex relaxations of OPF suffer from two major drawbacks: 1) Their global guarantees often come at the expense of higher runtimes or overly complicated implementations; 2) They do not account for the time-varying nature of demand. This time-varying property poses additional constraints on the ramping capabilities of generators, which in turn gives rise to coupled optimization problems.

On the other hand, research on multiperiod OPF, such as [12], [13], and dyanamic OPF, such as [14], [15], endeavors to solve multiple such time-coupled OPF problems simultaneously. This leads to large problem formulations which are still nonconvex in nature. As a result, solution strategies for these problems often rely on the convex relaxations discussed previously in combination with receding horizon approaches or nonlinear programming algorithms, which lack global optimality guarantees [12]. Another drawback is that the data for all time periods must be specified at the outset. In practice, forecasts may not be adequately accurate far in advance.

Real-time OPF is another approach which targets the timescale of OPF. In [16] a real-time algorithm is used to track the optimal solution every few seconds in between traditional OPF updates, which occur on a slower timescale ranging from every 5 to 30 minutes. It uses new measurements of the decision variables' values and constraints at every time step in order to compute a correction and track the optimal solution. The correction is computed by solving a quadratic optimization problem with one iteration of a quasi-Newton algorithm. This has the advantage of responding quickly to fluctuations, but does not replace the need to solve OPF on the traditional timescale. Other faster-timescale approaches to OPF-related problems include [17]–[19].

In this work, which is positioned between MPOPF and real-time OPF, we consider time-varying OPF with ramping constraints in an online fashion, where the load profile changes over time. Unlike the previous convexification techniques, we solve the problem sequentially using a simple local-search algorithm. Due to the nonconvex nature of the problem, the local-search algorithm may return a spurious (non-global) local solution, thus leading to a potentially large optimality gap. Previously in [20], we made the observation that for a small system with time-varying demand, the solution trajectories of the time-varying OPF stemming from four initial local solutions could converge over time. Here, we present an extensive empirical study on a larger system with 16 spurious solutions using California load data, and show that all feasible local solution sequences (also called trajectories) converge in cost and value to the best solution. Notably, this phenomenon occurs despite the fact that the problem has multiple point-wise poor local minima at key times. For this system, we show that there is an *escaping period* in which different local solution trajectories converge to a solution with lowest cost, followed by a *tracking period* in which the local trajectories closely track the global solution.

This observation leads to an important phenomenon in time-varying OPF: *load variation enables the local solution trajectories to avoid poor solutions over time*.[1] In other words, despite the highly nonconvex nature of the OPF problem at any given time, our numerical algorithm acts on an implicitly smoothed and well-behaved variant of the problem, thereby avoiding the undesirable local solutions over time. We will formalize this statement in the paper by providing a backward-in-time mapping from the globally optimal solutions of OPF at a given time (namely, end of the escaping period) to the set of desirable initial points. By leveraging its special structure, we show that the proposed backward mapping may act as a stochastic gradient ascent algorithm on an implicitly convexified formulation of the OPF problem, which in turn explains why local solution trajectories could avoid poor solutions over time. This work is the first studying the role of data variation in reducing the complexity of power optimization problems. Since it relies on simple local search methods, the solution techniques have extremely low memory and time complexities and can also be implemented in a distributed setting to accommodate the distributed nature of future grids [21].

[1] Note that with constant (time-invariant) load, all the local solution trajectories will remain unchanged over time.

II. Empirical Study of Time-Varying OPF

In this section, we analyze the local solution trajectories of time-varying OPF primarily for a 39-bus system. A secondary analysis on a 9-bus system is also shared to highlight that the observed behavior is not unique to the 39-bus system. The solution trajectories of time-varying OPF are constructed by sequentially solving a series of optimization problems with time-varying demand levels using a local-search algorithm. California load data and synthetic load scenarios are used to determine demand levels over time. To prevent the solution from changing abruptly over a short period of time, the sequential optimization problems are coupled via so-called *ramping constraints*, as we explain below.

A. Model and Algorithm Details

To examine the behavior of different local solution trajectories, we consider a modified version of the IEEE 39-bus system, as introduced in [4]. Specifically, the real and reactive power demands are reduced by 50%, voltage limits tightened from +/-6% to +/-5%, and the cost functions associated with all generators are assumed to be linear. The OPF problem for this system with a generation cost-minimizing objective and fixed demand values is known to have 16 local solutions. In this work, we take into account the time-varying nature of the load and scale all demands proportionally to a given load profile. Finally, we introduce the ramping constraints that limit the change in power generation for each generator over time.

Starting from the 16 known initial local solutions, we constructed the sequences of local trajectories using the MATPOWER optimization toolbox [22] and `fmincon` sequential quadratic programming solver[2] in the following procedure. We ran Algorithm 1 for all 16 initial local solutions and obtained 16 different solution sequences, which are called *discrete local trajectories* [20].

[2] Note that unlike many interior point methods that require strictly feasible initial points, `fmincon` sequential quadratic programming gives a second-order critical point even if the initial point is not strictly feasible.

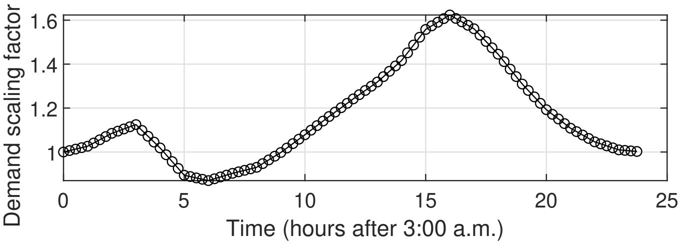


Fig. 1: Average daily net load for California during August 2019 [23]

Algorithm 1 Algorithm for obtaining discrete local trajectories

```
Input: Power system model with a fixed initial point $\mathbf{x}_0$, demand curve, ramping constraint specifications
Output: Discrete local trajectory $\{\mathbf{x}_t\}_{t=0}^{K}$
Initialization : $t = 1$
for every 15-minute time increment over 24 hours do
    Set demand constraints for each bus according to the demand curve at time $t$.
    Set generator production limits based on $\mathbf{x}_{t-1}$ and the ramping constraint.
    Solve the resulting cost-minimization OPF problem with fixed demand and initial point $\mathbf{x}_{t-1}$ using
    fmincon. Upon feasibility, collect the solution as $\mathbf{x}_t$.
end for
return $\{\mathbf{x}_t\}_{t=0}^{T}$
```

B. Behavior of Discrete Local Trajectories for a 39-bus system with California Data

In this example, the shape of the demand curve is based on the California's net load for an average day in August 2019 [23] (Fig. 1). The reported actual hourly net load data was interpolated linearly to produce a net load estimate for each 15-minute interval within 24 hours. The curve is normalized and shifted so that time 0 represents 3:00 a.m. Here, the maximum magnitude of allowable change in power generation between two consecutive time steps is 10% of the capacity of each generator. All 16 discrete local trajectories remain feasible throughout the span of twenty-four hours. (This is not guaranteed, as local search may not always find a feasible point or such point may not even exist.) Fig. 2 shows the point-wise distance between these feasible trajectories and the feasible trajectory with the lowest cost (labeled as *Trajectory 2*). Interestingly, all 16 trajectories converge to *Trajectory 2* within nine hours.

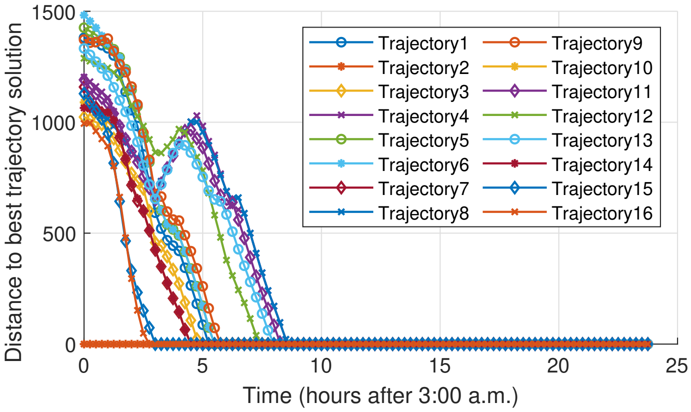


Fig. 2: Solution convergence for points on discrete local trajectories

Based on this observation, one may speculate that the problem becomes devoid of spurious local solutions over time. This is not the case for the considered problem. We uniformly searched the feasible region of the problem without ramping constraints and verified that there are multiple point-wise spurious local solutions for the point-wise (single time instance, without ramping constraints) OPF problem at different times. In particular, there are many local solutions around the *escape time* (hour 9) when the discrete local trajectories merge into one trajectory. Fig. 3 shows the normalized objective cost values for different discrete local trajectories, alongside the costs of the discovered point-wise local solutions. Despite the existence of multiple sub-optimal operating points at different times, the discrete local trajectories initialized at various local solutions result in the lowest cost values over time. Fig. 4 examines the active and reactive power generation for two representative generators. This figure shows that the problem has point-wise local solutions with a wide range of generation levels, highlighting the importance of finding the solution with the lowest cost.

Observe that most of the spurious point-wise local solutions have *sharp* and random nature. In other words, they appear at different time-steps with various cost values, and then quickly disappear after a short period of time. This implies that the landscape of OPF may be highly nonconvex at any given time step. However, it can be observed that our numerical algorithm is not affected by such sharp local solutions. To explain this phenomena, we will show in Section IV that the data variation enables the solver to act on a *smoothed version* of the problem that is devoid of sharp local minima.

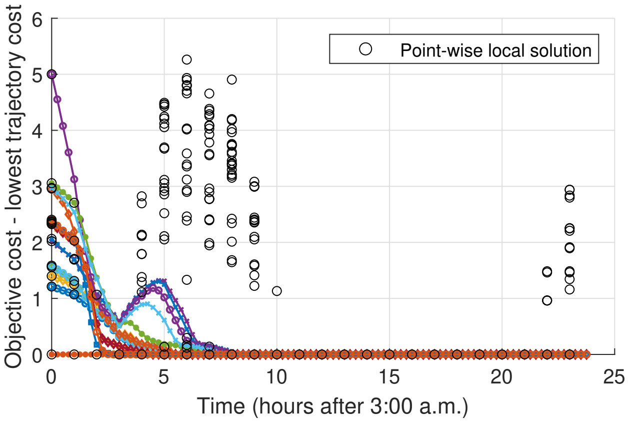


Fig. 3: Cost for points on discrete local trajectories and point-wise local solutions (for a single instance of OPF), relative to the cost of the best trajectory

C. Behavior of Discrete Local Trajectories for a 9-bus system with California Data

In this example, we consider a modified version of the IEEE 9-bus system with 4 known local solutions to the OPF problem, as introduced in [4]. Specifically, the active and reactive power demands

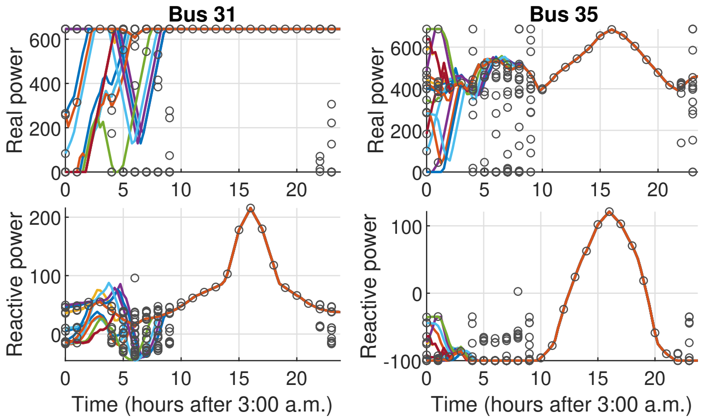
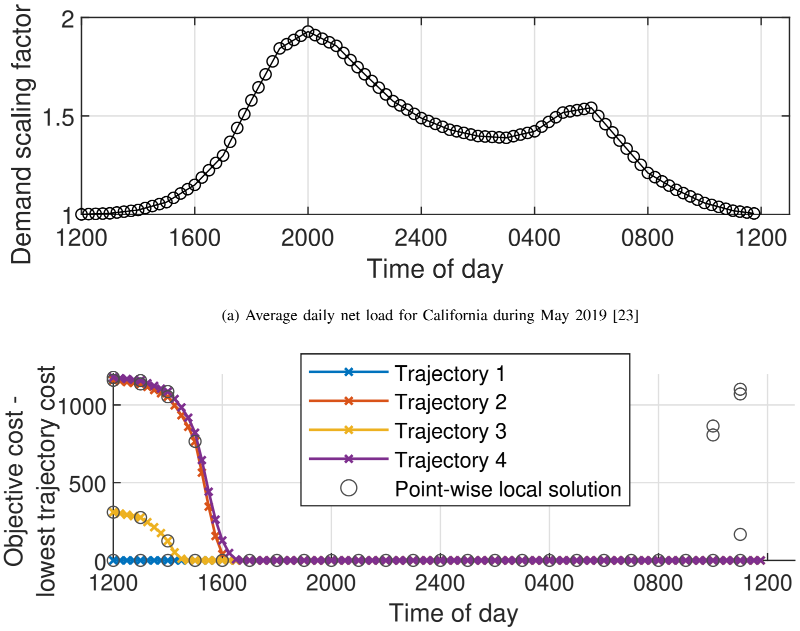


Fig. 4: Real and reactive power output of select generators: points on discrete local trajectories and point-wise local solutions



(a) Average daily net load for California during May 2019 [23]

(b) Cost for points on discrete local trajectories and point-wise local solutions (for a single instance of OPF), relative to the cost of the best trajectory

Fig. 5: Data and results for an empirical study on a 9-bus system

are reduced by 40% and the lower bounds on reactive power compensation are tightened to -5 Mvar. The demand data is a normalized and shifted version of California's net load for an average day in May 2019 [23] (Fig. 5a). Fig. 5b shows the relative objective cost values for different discrete local trajectories, alongside the costs of the discovered point-wise local solutions, produced using Algorithm 1 with a 5% ramping constraint. Again, we observe that the load variation enables all trajectories to converge to the optimal trajectory.

D. Impact of Load Variation on 39-bus System

Next, we consider discrete local trajectories for three different load profiles on the same 39-bus system. Isolating the impact of load variation enables insight into how variation creates trajectories that avoid poor solutions, as occurred in the previous examples. The three demand curves used are sinusoidal functions with amplitudes representing 5%, 10% and 12% deviation from the initial load, as shown in the left column of Fig. 6. The ramping constraint (i.e., maximum magnitude of allowable change in power generation between two consecutive time steps) is 5% of the capacity of each generator. In each scenario, all 16 discrete local trajectories remain feasible throughout the time horizon (100 steps).

The results show that larger magnitudes of data variation lead to fewer poor solutions over time. At 5% variation 4 trajectories remain at 4 different poor solutions, while the remaining 12 trajectories converge to the best solution. At 10% variation 3 trajectories converge to the same poor solution, while the remaining 13 trajectories converge to the best solution. At 12% variation all 16 trajectories converge to the best known solution. These results are displayed in the center column of Fig. 6, which shows the distance between each trajectory and the trajectory with the lowest cost, along with discovered point-wise local solutions. The search for point-wise local solutions is done every fourth time step due to the significant computational effort required to repeatedly solve the problem from a range of initial points. Fig. 6 (right column) compares the number of point-wise local solutions with the number of distinct[3] trajectories over time. In these three cases, *the number of distinct trajectories decreases until it plateaus at the minimum number of point-wise local solutions found over the entire period.* This offers one potential explanation of how load variation creates trajectories that escape poor solutions: In exploring a range of static problems, you may encounter one or more times at which the problem has a favorable landscape[4]. At such times, the coupled problem may escape a poor solution. Eventually, the number of poor trajectories is limited by the number of spurious point-wise local solutions of the most favorable landscape.

III. Mathematical Analysis of Time-Varying OPF

The case study in Section II reveals an important property of the time-varying OPF problem: In the *escaping period*, different discrete local trajectories converge to the operating point with the lowest cost. Then, in the *tracking period*, the discrete local trajectories track these globally optimal operating

[3] Solutions are considered distinct if the real or reactive power output at any generator differs by at least 1 MW or 1 MVAr, respectively, or if the voltage magnitude or angle at any bus differs by at least 10^{-3} p.u. (345V) or 10^{-3} radians, respectively.

[4] The number of spurious point-wise local solutions is an indicator of how difficult a given static OPF problem is. If only one point-wise local solution is found, the problem may be convex. However, the search is not exhaustive, so other local minima with small regions of attraction may exist.

points, even if the load profile changes gradually over time. Such *tracking period* has been studied in [24], [25], but the striking feature of power systems is the existence of escaping periods.

To better understand this phenomenon, we analyze the problem structure mathematically. First, we reformulate the time-varying OPF as an unconstrained optimization problem to enable the analysis. Using the derived unconstrained problem, we introduce a backward mapping that characterizes the dynamics of the discrete local trajectories over time. We show that the convergence of different local trajectories can be explained by the expansive property of this backward mapping. Finally, in Section IV we draw a novel connection between our derived mapping and stochastic gradient ascent and use this insight to explain that the behavior of the trajectories may be driven by some low-complexity averaged model over a period, rather than the high-complexity OPF problems at each step.

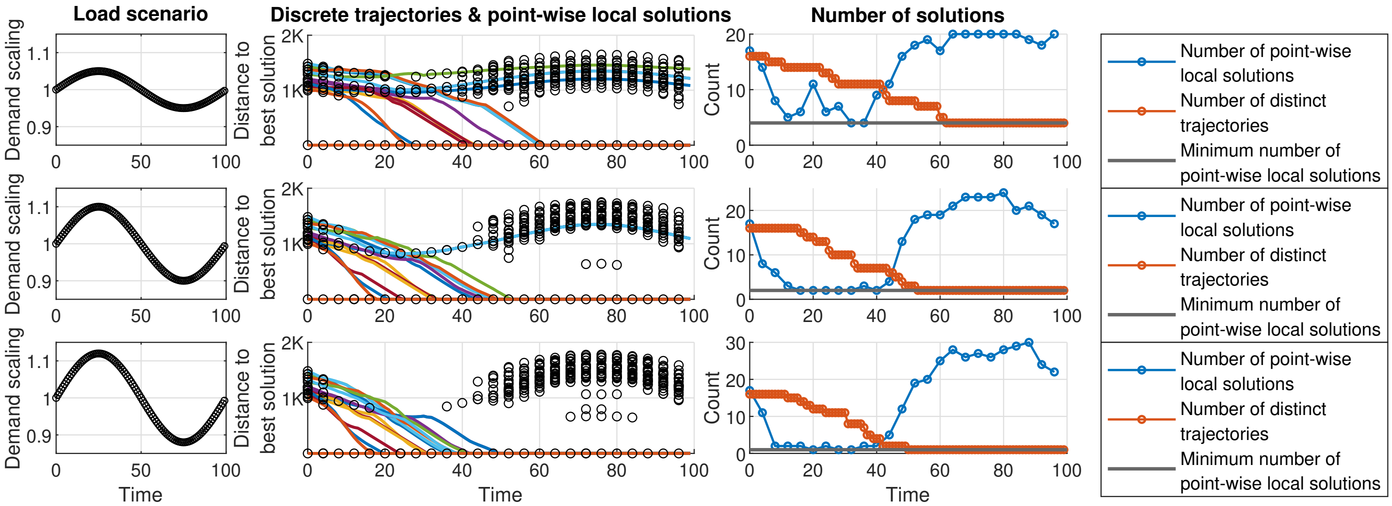


Fig. 6: Three scenarios highlighting the role of load variation. The three plots for each scenario are (left to right): load profile input, resulting discrete local trajectories and point-wise local solutions (for a single instance of OPF), comparison of the number of point-wise local solutions to the number of distinct discrete local trajectories.

A. Unconstrained Model for OPF with Fixed Demand

The AC model of OPF in a single time instance with fixed and predefined demand values can be written compactly as an optimization problem with both equality and inequality constraints:

$$\min_{\mathbf{x}\in\mathbb{R}^p} \ f(\mathbf{x}) \tag{1a}$$

$$\text{s.t. } h(\mathbf{x}) = \mathbf{d} \in \mathbb{R}^n \tag{1b}$$

$$g(\mathbf{x}) \leq \mathbf{0} \in \mathbb{R}^m \tag{1c}$$

Here, $\mathbf{x}$ is the concatenation of the voltage angle and magnitude at each bus, as well as the real and reactive power generation outputs for each generator. The equality constraint (1b) ensures that the generated power meets the demand, where $\mathbf{d}$ is the vector of real and reactive demand at each bus, and respects the underlying structure and physical constraints of the network. The remaining constraints in the problem—including the upper and lower bounds on the voltage magnitudes and degrees, power generation, and line flows—are captured by the inequality constraint (1c). It is easy to verify that $p > n$. We refer the reader to [2], [4] and [9] for more information on the exact formulation of the problem. Note that $f(\mathbf{x})$, $h(\mathbf{x})$, and $g(\mathbf{x})$ are continuously differentiable (piecewise linear cost functions can be reformulated as such).

In order to analyze this optimization problem theoretically, it is desirable to convert it to an unconstrained problem. First, we enforce the inequality constraints (1c) through a penalty in the objective function:

$$\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + \beta \sum_{i=1}^{m} \left([g_i(\mathbf{x})]^+\right)^2 \tag{2a}$$

$$\text{s.t. } h(\mathbf{x}) = \mathbf{d} \in \mathbb{R}^n \tag{2b}$$

where $\beta > 0$ serves as the penalization parameter, $g_i(\cdot)$ is the i^{th} element of $g(\cdot)$, and $[y]^+$ denotes $\max(y, 0)$. This choice of quadratic penalty function is inexact, meaning that problem (2) is an approximation of problem (1). However, as β increases, each global minimizer of (1) approaches a global minimizer of (2) under mild regularity conditions [26]. Second, we use the implicit function theorem [26] to complete the transformation to an unconstrained model. Consider a feasible point $\mathbf{x}_\star$ satisfying the Karush-Kuhn-Tucker (KKT) conditions for (2). Assuming that constraint qualifications hold at $\mathbf{x}_\star$, this vector can be partitioned into two sub-vectors ${\mathbf{x}_\star}^B \in \mathbb{R}^n$ and ${\mathbf{x}_\star}^R \in \mathbb{R}^{p-n}$ such that the Jacobian of $h(\mathbf{x}_\star)$ with respect to $\mathbf{x}^B$ is invertible. Therefore, the implicit function theorem guarantees the existence of a unique differentiable function $\phi(\cdot)$ such that $\mathbf{x}^B = \phi(\mathbf{x}^R)$ in a local neighborhood of $\mathbf{x}_\star$. Given such function, Problem (2) can be re-written as (see [26]):

$$\min_{\mathbf{x}^R\in\mathbb{R}^{p-n}} f(\phi(\mathbf{x}^R), \mathbf{x}^R) + \beta \sum_{i=1}^{m} \left([g_i(\phi(\mathbf{x}^R), \mathbf{x}^R)]^+\right)^2 \tag{3}$$

Enforcing the equality constraint (1b) directly using the implicit function theorem instead of through penalization will be advantageous when we move to the time-varying setting. Namely, it avoids amplifying the demand variation as scaling by a large penalization parameter would do. This is not an issue for the inequality constraint (1c) because it does not vary in time.

Remark 1: Note that (3) *cannot be formulated explicitly, due to the unknown nature of the local solution* $\mathbf{x}_\star$ *and the function* $\phi(\mathbf{x}^R)$. *Instead, this formulation serves as an intermediate step to analyze* *the behavior of discrete local trajectories over time. In other words, one would solve the OPF problem directly in practice, and the surrogate problem* (3) *is designed to understand the properties of OPF.*

B. Unconstrained Model for Time-Varying OPF

The above analysis reveals that, under some technical conditions, the OPF problem with fixed load can be modeled as an unconstrained optimization problem (with a controllable approximation error). In this subsection, we extend our analysis to time-varying OPF where demand changes over time and the problem must respect ramping constraints. As previously stated, ramping constraints ensure that the solution does not change too drastically from one time step to the next. One way to softly impose ramping constraints is through a proximal method, which penalizes the distance between the current and previous solutions in the objective function of the optimization [27]. Time-varying OPF with K equally-spaced time steps $t_0 = 0, t_1 = \Delta t, ..., t_K = K\Delta t$ ($\Delta t > 0$) can be written as the following sequence of optimization problems:

$$\min_{\mathbf{x}^{R_k} \in \mathbb{R}^{p-n}} f_{t_k}(\phi_{t_k}(\mathbf{x}^{R_k}), \mathbf{x}^{R_k}) + \alpha \left\| \mathbf{x}^{R_k} - \mathbf{x}_{\star t_{k-1}}^{R_k} \right\|_2^2 + \beta \sum_{i=1}^{m} ([g_i\left(\phi_{t_k}(\mathbf{x}^{R_k}), \mathbf{x}^{R_k}\right)]^+)^2 \tag{4}$$

for $k = 1, ..., K$, where $\alpha > 0$ is a penalization parameter and $\mathbf{x}_{\star t_{k-1}} = \left[\left(\mathbf{x}_{\star t_{k-1}}^{B_k}\right)^\top \quad \left(\mathbf{x}_{\star t_{k-1}}^{R_k}\right)^\top \right]^\top$ denotes an arbitrary local solution to Problem (4) obtained at time t_{k-1}. In light of its dependence on $\mathbf{x}^{R_k}$, $\mathbf{x}^{B_k}$ is not regularized in this approximated model. Due to the time-varying nature of the demand, the functions $f_{t_k}(\cdot)$ and $\phi_{t_k}(\cdot)$ may change over time, hence they are indexed by time step.

To simplify the analysis, assume that the partition (B_k, R_k) does not change over time, i.e., we have $B_k = B$ and $R_k = R$ for $k = 1, ..., K$. Then problem (4) can be written as

$$\min_{\mathbf{z} \in \mathbb{R}^{p-n}} \underbrace{F_k(\mathbf{z}) + \beta \sum_{i=1}^{m} \left([G_{k,i}(\mathbf{z})]^+\right)^2}_{\Gamma_k(\mathbf{z})} + \alpha \left\| \mathbf{z} - \mathbf{z}_{k-1} \right\|_2^2 \tag{5}$$

for $k = 1, ..., K$, where $\mathbf{z} = \mathbf{x}^{R_k}$, $\mathbf{z}_{k-1} = \mathbf{x}_{\star t_{k-1}}^{R_k}$, $F_k(\mathbf{z}) := f_{t_k}(\phi_{t_k}(\mathbf{z}), \mathbf{z})$, and $G_{k,i}(\mathbf{z}) := g_i\left(\phi_{t_k}(\mathbf{z}), \mathbf{z}\right)$. If the partition changes, then the time interval $[0, K\Delta t]$ should be divided into sub-intervals, each with a constant partitioning of $\mathbf{x}$. In this case, the argument presented in Section IV applies to each sub-interval.

C. Backward-In-Time Mapping

The above analysis reveals that a local-search algorithm used to solve the time-varying OPF implicitly aims to recover a stationary point of the unconstrained problem (5). Therefore, we focus on (5) in our subsequent analysis. Consider a given time step $T\Delta t$, representing the end of the escaping period. Then, a sequence of stationary points $\{\mathbf{z}_k\}_{k=1}^T$ for (5) satisfies

$$0 = \nabla \Gamma_k(\mathbf{z}_k) + 2\alpha(\mathbf{z}_k - \mathbf{z}_{k-1}) \tag{6}$$

for every $k = 1, 2, \ldots, T$ (where ∇ is the gradient operator). Note that $\Gamma_k(\cdot)$ is differentiable. Therefore, given the solution $\mathbf{z}_{k-1}$, this equation defines an implicit nonlinear formula for obtaining $\mathbf{z}_k$ which cannot be written in closed form. However, going backward in time, one can express $\mathbf{z}_{k-1}$ in terms of $\mathbf{z}_k$:

$$\mathbf{z}_{k-1} = \mathbf{z}_k + \frac{1}{2\alpha}\nabla \Gamma_k(\mathbf{z}_k) := M_k(\mathbf{z}_k) \tag{7}$$

This gives rise to the following end-to-end backward mapping from $\mathbf{z}_T$ to the initial point $\mathbf{z}_0$ via the composition operator $\circ$:

$$\mathbf{z}_0 = M_1 \circ M_2 \circ \cdots \circ M_T(\mathbf{z}_T) \tag{8}$$

Provided that the mapping $M_1 \circ \cdots \circ M_T(\mathbf{z}_T)$ is expansive enough when applied to a small neighborhood of a global solution of OPF at time $T\Delta t$, a large set of initial points (even multiple local solutions of OPF at time 0) are guaranteed to converge to that small neighborhood of the globally optimal solution of the problem at time $T\Delta t$. This expansive nature of the mapping implies the escape of spurious local solutions between time 0 and time $T\Delta t$. The global solutions at future times after $T\Delta t$ will be tracked successfully if the data variation is not too high [24]. This expansive property can be observed in the empirical study conducted in Section II on the modified IEEE 39-bus and 9-bus system under both California load data and synthetic sinusoidal loads.

IV. Connection to Stochastic Gradient Ascent

This section aims to explain how data variation plays a key role in escaping spurious local solutions of time-varying OPF. Specifically, we will show that the backward mapping (7) can be treated as a variant of stochastic gradient ascent on a smoothed version of the function $\Gamma_T(\mathbf{z})$. This gives rise to the following important observation:

A certain level of stochasticity in $\{\Gamma_k(\mathbf{z})\}_{k=1}^T$ over time may enable the stationary points $\{\mathbf{z}_k\}_{k=1}^T$ to escape "sharp" local minima over time.

To explain this phenomenon, we first introduce the smoothing property of the stochastic gradient descent (SGD) algorithm.

Smoothing property of SGD: Recently, [28] proposed an alternative viewpoint to SGD and its ability to avoid spurious sharp local minima. Given an initial point $\mathbf{z}_0$, suppose our goal is to find a global minimum of a (time-invariant) function $\Gamma(\mathbf{z})$ using SGD. Accordingly, the iterations of SGD can be written

$$\mathbf{z}_{k+1} = \mathbf{z}_k - \eta(\nabla\Gamma(\mathbf{z}_k) + \omega_k) \qquad \forall k \in \{0, 1, 2, \dots\} \tag{9}$$

where ω_t is a bounded random variable with zero mean and η is a predefined step size. Upon defining $\tilde{\mathbf{z}}_k = \mathbf{z}_k - \eta\nabla\Gamma(\mathbf{z}_k)$, one can write the above iterations (9) in terms of the intermediate sequence $\{\tilde{\mathbf{z}}_k\}$:

$$\tilde{\mathbf{z}}_{k+1} = \tilde{\mathbf{z}}_k - \eta\omega_k - \eta\nabla\Gamma(\tilde{\mathbf{z}}_k - \eta\omega_k), \forall k \in \{0, 1, 2, \dots\} \tag{10}$$

To analyze the average behavior of SGD, consider the evolution of $\mathbb{E}_{\omega_k}(\tilde{\mathbf{z}}_{k+1})$, where the expectation is taken over ω_k conditioned on $\{\omega_0, \dots, \omega_{k-1}\}$. Hence,

$$\mathbb{E}_{\omega_k}[\tilde{\mathbf{z}}_{k+1}] = \tilde{\mathbf{z}}_k - \eta\nabla\mathbb{E}_{\omega_k}[\Gamma(\tilde{\mathbf{z}}_k - \eta\omega_k)], \forall k \in \{0,1,2,..\} \tag{11}$$

Therefore, on average, SGD acts as the exact gradient descent on the surrogate function $\mathbb{E}_{\omega_k}[\Gamma(\tilde{\mathbf{z}}_k - \eta\omega_k)]$. Comparing this function with $\Gamma(\mathbf{z})$, one can verify that the former is a smoothed version of the latter, where the smoothness is due to the convolution of $\Gamma(\mathbf{z})$ with the probability density function of the random variable ω_k. As illustrated in [28], such convolution may give rise to (one-point) strong convexity of $\mathbb{E}_{\omega_k}[\Gamma(\tilde{\mathbf{z}}_k - \eta\omega_k)]$ with respect to the globally optimal solution, which in turn guarantees the convergence of $\{\tilde{\mathbf{z}}_k\}$ (and hence $\{\mathbf{z}_k\}$) to a small neighborhood around the global solution, even in the presence of sharp local minima. A key takeaway from this observation is that $\Gamma(\mathbf{z})$ can possess multiple sharp, poor local minima, and yet its smoothed version $\mathbb{E}_{\omega_k}[\Gamma(\tilde{\mathbf{z}}_k - \eta\omega_k)]$ may be devoid of such solutions.

Time-varying optimization and time-varying OPF: Returning to time-varying OPF and the backward mapping (7), we assume that the variation in $\{\nabla\Gamma_k(\mathbf{z})\}_{k=1}^T$ follows a stochastic process indexed by the time k. In particular, we write $\nabla\Gamma_k(\mathbf{z}) - \nabla\Gamma_{k+1}(\mathbf{z}) = \zeta_k(\mathbf{z}) + \omega_k$, where $\zeta_k(\mathbf{z})$ is a deterministic, time-varying function and ω_k is a bounded random variable with zero mean. Such assumption is realistic in power systems, where demand can be modeled as a deterministic, time-varying function capturing the average demand behavior, together with an additive stochastic term accounting for its random nature. The iteration (7) is equivalent to

$$\begin{aligned}\mathbf{z}_k =& \mathbf{z}_{k+1} + \frac{1}{2\alpha}\nabla\Gamma_T(\mathbf{z}_{k+1}) \\ &+ \frac{1}{2\alpha}\sum_{\tau=k+1}^{T-1}\underbrace{(\nabla\Gamma_\tau(\mathbf{z}_{k+1}) - \nabla\Gamma_{\tau+1}(\mathbf{z}_{k+1}))}_{\zeta_\tau(\mathbf{z}_{k+1})-\omega_\tau}\end{aligned} \tag{12}$$

which can be written as the following dynamical model:

$$\mathbf{z}_k = \mathbf{z}_{k+1} + \frac{1}{2\alpha}\nabla\Gamma_T(\mathbf{z}_{k+1}) + \frac{1}{2\alpha}\nu_{k+1}(\mathbf{z}_{k+1}) \tag{13a}$$

$$\nu_{k+1}(\mathbf{z}_{k+1}) = \nu_{k+2}(\mathbf{z}_{k+1}) + \zeta_{k+1}(\mathbf{z}_{k+1}) - \omega_{k+1} \tag{13b}$$

where $\nu_{k+1}(\mathbf{z}_{k+1})$ is referred to as the *variation process*. In particular, (13b) defines explicit dynamics for the variation process comprised of three parts. The first term $\nu_{k+2}(\mathbf{z}_{k+1})$ captures the correlation between the variation processes at times t_{k+1} and t_{k+2}. The second term $\zeta_{k+1}(\mathbf{z}_{k+1})$ captures the *bias* that is added to the variation process at time t_{k+1}. Lastly, the third term $\omega_{k+1} \sim W(\mathbf{z}_{k+1})$ is an independent noise injected into the variation process at time t_{k+1} (also referred to as *effective noise*). Comparing (13) with (9), one can verify that (13) reduces to stochastic gradient ascent if $\nu_{k+2}(\mathbf{z}_{k+1}) + \zeta_{k+1}(\mathbf{z}_{k+1}) = 0$. Therefore, if ω_{k+1} dominates the first two terms, (13) resembles an approximate version of stochastic gradient ascent applied to $\Gamma_T(\mathbf{z})$; otherwise, it is a *biased* and *correlated* version of SGD [29]. Similar to (11), this implies that, on average, the points generated via the backward mapping (7) would be close to the iterations of the gradient ascent on the smoothed version of $\Gamma_T(\mathbf{z})$. Now, assume that despite the possible existence of multiple spurious and sharp local minima in $\{\Gamma_k(\mathbf{z})\}_{k=1}^T$, the smoothed version of $\Gamma_T(\mathbf{z})$ after convolution is strongly convex. This together with the expansive nature of gradient ascent on strongly convex functions [30] yields that the end-to-end backward mapping (8) is expansive, and the discrete local trajectories can escape poor local solutions over time. We formalize and rigorously analyze this intuition in the next subsection.

A. Theoretical analysis of dynamics

For simplicity of notation, we define $\eta = \frac{1}{2\alpha}$. Furthermore, suppose that $\mathbf{z}^*$ denotes the globally minimum point of $\Gamma_T(\mathbf{z})$. Without loss of generality, $\|\mathbf{v}\|$ is used to refer to the 2-norm of the vector $\mathbf{v}$. We make the following assumptions for the dynamical model (13):

Assumption 1 (Smoothness): The following statements hold:

- The function $\Gamma_T(\mathbf{z})$ is L-smooth, i.e., we have

$$\|\nabla\Gamma_T(\mathbf{x}) - \nabla\Gamma_T(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\| \ \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^{p-n}. \tag{14}$$

- The functions $\zeta_\tau(\mathbf{z})$ are l-Lipschitz for $\tau = 1, \cdots, T-1$, i.e., we have

$$\|\zeta_k(\mathbf{x}) - \zeta_k(\mathbf{y})\| \le l\|\mathbf{x}-\mathbf{y}\| \ \forall \mathbf{x},\mathbf{y} \in \mathbb{R}^{p-n}. \tag{15}$$

Assumption 2 (Implicit Convexity): There exists $\mathbf{z}^*$ such that the following statements hold:

- (One-point strong convexity of convolution) For every $\mathbf{y}$, there exists $c > 0$ such that

$$\langle \mathbf{z}^* - \mathbf{y}, -\nabla\mathbb{E}_{\omega\sim W(\mathbf{z})}\left[\Gamma_T(\mathbf{y}-\eta\omega)\right]\rangle \ge c\|\mathbf{y}-\mathbf{z}^*\|^2 \tag{16}$$

- (Bounded one-point curvature of convolution) For every $\mathbf{y}$, there exists $c' > 0$ such that

$$\langle \mathbf{z}^* - \mathbf{y}, -\sum_{\tau=k+1}^{T-1}\mathbb{E}_{\omega\sim W(\mathbf{z})}\left[\zeta_\tau(\mathbf{y}-\eta\omega)\right]\rangle \ge -c'\|\mathbf{y}-\mathbf{z}^*\|^2 \tag{17}$$

for every $k \in \{0, \dots, T-2\}$.

The existence of L and l which satisfy Assumption 1 can be verified for the unconstrained model of the time-varying OPF. Meanwhile, Assumption 2 implies that the *convoluted* variant of the objective function at time T is one-point strongly convex. We note that such assumption may not be easily verifiable for the time-varying OPF. However, our simulations strongly support the fact that most of the spurious local solutions in time-varying OPF have a sharp nature, and therefore, they are likely to be absent in the convoluted (smoothed) landscape of the problem.

Under these two assumptions, we present the main theorem of this paper.

Theorem 1: Suppose that $c \geq c'$ and there exists $r \geq 1$ such that $\|\omega_t\| \leq r$ for every t. Define $\lambda := \eta(c - c')$, and assume that $2\eta^2 L < 1$. Then, under Assumptions 1 and 2, the following inequality holds:

$$\|\mathbf{z}_T - \mathbf{z}^*\|^2 \leq \frac{1}{1-2\eta^2 L}\left(\mathcal{D} + \frac{\mathbb{E}\left[\|\mathbf{z}_0 - \mathbf{z}^*\|^2\right]}{(1+\lambda)^{T-1}} + \frac{8\eta^2 r^2 T^2}{(1+\lambda)^{T-1}}\right) \tag{18}$$

where

$$\mathcal{D} = \left(\frac{4}{\lambda} + \frac{4}{\lambda^2}\right)\eta^3 r^2 l + 16\left(1 + \frac{1}{\lambda}\right)^2 \frac{\eta^2 r^2 (1+2\lambda)^2}{\lambda^2} \tag{19}$$

A sketch of the proof for Theorem 1 is provided in the appendix. A number of observations can be made based on this theorem. Not surprisingly, the provided bound on $\|\mathbf{z}_T - \mathbf{z}^*\|$ depends on the *accuracy of the initial point* $\|\mathbf{z}_0 - \mathbf{z}^*\|$. However, the effect of this initial accuracy diminishes exponentially fast with respect to T. Moreover, as $T \to \infty$, the following asymptotic inequality holds:

$$\|\mathbf{z}_T - \mathbf{z}^*\|^2 \leq \frac{\mathcal{D}}{1 - 2\eta^2 L} \tag{20}$$

which is independent of the initial point. Another implication of this asymptotic bound is that, for any value of T, Theorem 1 can only guarantee the convergence of $\mathbf{z}_T$ to a neighborhood of $\mathbf{z}^*$. This is not surprising if we consider the non-diminishing nature of η and its connection to SGD, as delineated in the introduction of Section IV. Indeed, similar results on SGD show that, with non-diminishing step-sizes, the iterations of the algorithm may only converge to a neighborhood of the globally optimal solution [28]. Finally, it is worthwhile to study how $\mathcal{D}$ depends on different parameters of problem, namely η, r, l, L, and $c - c'$. Equation (19) reveals that $\mathcal{D}$ is a decreasing function of $c - c'$. Combined with Assumption 2, this implies that one-point strong convexity of $\Gamma_t(\mathbf{z})$ for $t = 1, \dots, T$ has a favorable effect on $\|\mathbf{z}_T - \mathbf{z}^*\|$. Similarly, it can be seen from (18) and (19) that $\|\mathbf{z}_T - \mathbf{z}^*\|$ decreases as l, L, and the noise values' magnitude (characterized by r) shrink. However, notice that Assumption 2 may not be satisfied for small values of noise. Finally, $\mathcal{D}$ does not have a monotone behavior with respect to η. In particular, it can

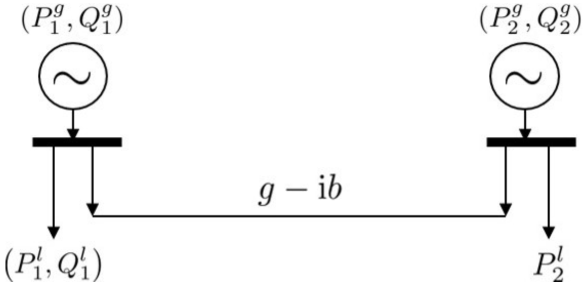


Fig. 7: The 2-bus system. Here, $\mathrm{i} = \sqrt{-1}$.

be verified that $\mathcal{D} \to \infty$ if $\eta \to \infty$ or $\eta \to 0^+$. Recalling (5) and $\eta = \frac{1}{2\alpha}$, this implies that over- or under-regularization may lead to large values for $\|\mathbf{z}_T - \mathbf{z}^*\|$. This observation is in line with Example 1 of [20], which shows that both small and large regularization may cause the solution trajectory to remain trapped at spurious local solutions of a time-varying optimization.

B. Illustrative Example on a 2-bus System

In this subsection, we analyze the effect of the load variation on the landscape of a 2-bus system. Our goal is to visualize the smoothing effect of the load variation on the objective function, thereby verifying the assumption on the implicit one-point strong convexity of the convoluted objective function. Consider the simple 2-bus system illustrated in Figure 7. Assume that both buses are equipped with generators, and they are connected via a single line with admittance $g - \mathrm{i}b$. The time-varying load connected to the first bus has both active and reactive power demands, while the time-varying load connected to the second bus is purely active. At any given time k, the point-wise OPF (without ramping constraints) can be formulated as follows[5]:

$$\min\ f_1(P_1^g) + f_2(P_2^g) \tag{21a}$$

$$\text{s.t. } P_1^g - P_{1;k}^l = |v_1|^2 g + |v_1||v_2| b \sin(\Delta\theta) - |v_1||v_2| g \cos(\Delta\theta) \tag{21b}$$

$$P_2^g - P_{2;k}^l = |v_2|^2 g + |v_1||v_2| b \sin(\Delta\theta) - |v_1||v_2| g \cos(\Delta\theta) \tag{21c}$$

$$Q_1^g - Q_{1;k}^l = |v_1|^2 g - |v_1||v_2| g \sin(\Delta\theta) - |v_1||v_2| b \cos(\Delta\theta) \tag{21d}$$

$$Q_2^g = |v_2|^2 g - |v_1||v_2| g \sin(\Delta\theta) - |v_1||v_2| b \cos(\Delta\theta) \tag{21e}$$

$$V^{\min} \le |v_1| \le V^{\max}, \quad V^{\min} \le |v_2| \le V^{\max} \tag{21f}$$

$$P_1^{\min} \le P_1^g \le P_1^{\max}, \quad P_2^{\min} \le P_2^g \le P_2^{\max} \tag{21g}$$

$$Q_1^{\min} \le Q_1^g \le Q_1^{\max}, \quad Q_2^{\min} \le Q_2^g \le Q_2^{\max} \tag{21h}$$

where P_i^g, Q_i^g, $|v_i|$, $\Delta\theta$ are the variables for active power generation, reactive power generation, voltage magnitude at bus i, and angle difference between buses 1 and 2 respectively. Moreover, $P_{i;k}^l$, $Q_{i;k}^l$ are the active and reactive load parameters at bus i and time k, respectively. To simplify our subsequent analysis, we assume that the voltage magnitudes at both buses are equal to the their nominal values, i.e., $|v_1| = |v_2| = 1$. Therefore, according to (21b)-(21e), the variables $(P_1^g, P_2^g, Q_1^g, Q_2^g)$ can be written in terms of the angle differences $\Delta\theta$. In other words, $P_1^g = p_1(\Delta\theta, P_{1;k}^l)$, $P_2^g = p_2(\Delta\theta, P_{2;k}^l)$, $Q_1^g = q_1(\Delta\theta, Q_{1;k}^l)$, $Q_2^g = q_2(\Delta\theta)$ where

$$p_1(\Delta\theta, P_{1;k}^l) = P_{1;k}^l + g + b\sin(\Delta\theta) - g\cos(\Delta\theta)$$

$$p_2(\Delta\theta, P_{2;k}^l) = P_{2;k}^l + g + b\sin(\Delta\theta) - g\cos(\Delta\theta)$$

$$q_1(\Delta\theta, Q_{1;k}^l) = Q_{1;k}^l + g - g\sin(\Delta\theta) - b\cos(\Delta\theta)$$

$$q_2(\Delta\theta) = g - g\sin(\Delta\theta) - b\cos(\Delta\theta)$$

[5] For simplicity, we omit the apparent power flow limits on the line connecting the two buses. Moreover, to streamline our subsequent analysis, we avoid the index k for the variables.

Based on these simplifications, the OPF at time k can be re-written as

$$\begin{aligned}\min\quad & f_1(p_1(\Delta\theta, P^l_{1;k})) + f_2(p_2(\Delta\theta, P^l_{2;k})) && \text{(22a)}\\ \text{s.t.}\quad & P_1^{\min} \leq p_1(\Delta\theta, P^l_{1;k}) \leq P_1^{\max}, && \text{(22b)}\\ & P_2^{\min} \leq p_2(\Delta\theta, P^l_{2;k}) \leq P_2^{\max} && \text{(22c)}\\ & Q_1^{\min} \leq q_1(\Delta\theta, Q^l_{1;k}) \leq Q_1^{\max}, && \text{(22d)}\\ & Q_2^{\min} \leq q_2(\Delta\theta, Q^l_{2;k}) \leq Q_2^{\max} && \text{(22e)}\end{aligned}$$

Moreover, suppose that the upper and lower bounds on the active and reactive power generations are chosen such that all inequality constraints in (22) remain inactive, except for lower bound on the reactive power generation, i.e., $Q_1^{\min} \leq q_1(\Delta\theta, Q^l_{1;k})$. Similar to (1), we convert (22) to an unconstrained optimization by removing this constraint, and instead, penalizing its violation in the objective function. Based on these modifications, we arrive at the following nonconvex and unconstrained optimization problem:

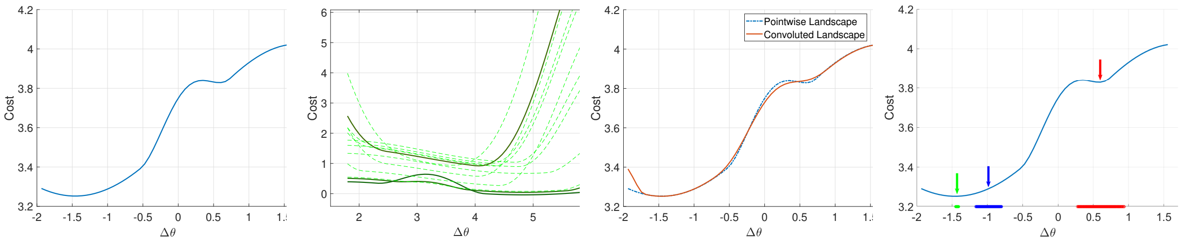
$$\begin{aligned}\min_{\Delta\theta}\quad \Gamma_k(\Delta\theta) =& f_1(p_1(\Delta\theta, P^l_{1;k})) + f_2(p_2(\Delta\theta, P^l_{2;k}))\\ &+ \beta\left(\left[Q_{\min} - q_1(\Delta\theta, Q^l_{1;k})\right]^+\right)^2 \qquad \text{(23)}\end{aligned}$$

Suppose that $g - \mathrm{i}b = 0.01 - \mathrm{i}0.1$ and $Q_{\min} = -0.181$. Moreover, suppose that $f_1(P_1^g) = 2(P_1^g)^2 + 2P_1^g + 1$ and $f_2(P_2^g) = 0.1(P_2^g)^2 + 0.1P_2^g + 1$. Finally, the penalization parameter β is set to 500. Figure 8a illustrates the objective function at the final time T as a function of $\Delta\theta$ for the choices of $P^l_{1;T} = P^l_{2;T} = 0.5$, and $Q^l_{1;T} = Q^l_{2;T} = 0$. Note that the objective function has one global minimum, one strict local minimum, and one local maximum within the interval $-2 \leq \Delta\theta \leq 1.5$.

Next, we illustrate the effect of load variation on the landscape of this optimization problem and verify Assumption 2. We empirically compute the function $\mathbb{E}_{\omega\sim W(\Delta\theta)}\left[\Gamma_T(\Delta\theta - \eta\omega)\right]$ introduced in Assumption 2 when the active and reactive loads are chosen according to the following rules:

- $P^l_{1;k}$ and $P^l_{2;k}$ are chosen uniformly at random from the interval $[0.005, 0.55]$.
- $Q^l_{2;k} = 0$ and $Q^l_{1;k}$ is chosen uniformly at random from the interval $[-0.02, 0.18]$.

Setting $\eta = 2$, for every $k = 0, 1, \dots, N = 10,000$ we randomly generate the active and reactive load values based on the aforementioned rules, and compute $\Gamma_k(\Delta\theta)$ and $\nabla\Gamma_k(\Delta\theta)$. Figure 8b shows realizations of $\Gamma_k(\Delta\theta)$ for different values of k. Then, for every $k = 0, 1, \dots, N-1$, we compute the gradient difference $\nabla\Gamma_k(\Delta\theta) - \nabla\Gamma_{k+1}(\Delta\theta)$, capturing the effects of the bias $\zeta_k(\Delta\theta)$ and the effective noise $w_k \sim W(\Delta\theta)$. Since the load distribution is the same at every time, we have $\mathbb{E}[\Gamma_k(\Delta\theta)] = \mathbb{E}[\Gamma_{k+1}(\Delta\theta)]$. Hence $\zeta_k(\Delta\theta) = 0$ for every k. Finally, we approximate $\mathbb{E}_{\omega\sim W(\Delta\theta)}\left[\Gamma_T(\Delta\theta - \eta\omega)\right]$ with its empirical



(a)

(b)

(c)

(d)

Fig. 8: (a) The objective function at $t = T$, (b) instances of the objective function for different values of the load, (c) the convoluted and pointwise objective functions, (d) realizations of $\Delta\theta - \eta\omega$ showing the effective noise of the load variation at different $\Delta\theta$ points.

counterpart $\frac{1}{N}\sum_{k=0}^{N-1}\Gamma_T(\Delta\theta - \eta\omega_k(\Delta\theta))$.[6] The resulting function for $-2 \leq \Delta\theta \leq 1.5$ is depicted in Figure 8c. It can be seen that, unlike $\Gamma_T(\Delta\theta)$, the convoluted function is devoid of spurious local minimum. In fact, it is one-point strongly convex, thereby verifying Assumption 2 on the implicit convexity of the convoluted objective function.

C. The effect of expected gradient on the noise variance

Another interpretation of the smoothing effect of the noise is based on the average behavior of the objective function. We will show that the variance of the effective noise $\mathbb{E}_{\omega\sim W(\Delta\theta)}[\|w\|^2]$ at a given point $\widehat{\Delta\theta}$ depends on the gradient of the *expected objective function* (where the expectation is taken over the randomness of the load). In other words, a large gradient of the expected objective function at $\widehat{\Delta\theta}$ leads to a high variance $\mathbb{E}_{\omega\sim W(\widehat{\Delta\theta})}[\|w\|^2]$, which in turn yields a smoother $\mathbb{E}_{\omega\sim W(\widehat{\Delta\theta})}\left[\Gamma_T(\widehat{\Delta\theta} - \eta\omega)\right]$. Figure 8d precisely shows this behavior. In particular, the local minimum $\Delta\theta = 0.6$ of $\Gamma_T(\Delta\theta)$ disappears in $\mathbb{E}_{\omega\sim W(\Delta\theta)}\left[\Gamma_T(\Delta\theta - \eta\omega)\right]$ due to the high variance of the additive noise ω at $\Delta\theta = 0.6$ (shown with red circles). On the other hand, the additive noise at the global minimum $\Delta\theta = -1.4$ is infinitesimal due to the fact that the gradient of the average function remains close to zero at $\Delta\theta = -1.4$. We will now formalize this intuition.

To better elucidate the relationship between the effective noise variance and the expected gradient of the objective function, consider an n-bus system with the following properties:

- Every bus i is equipped with a generator.
- The upper and lower bound constraints on the reactive power generations, and the upper bound constraints on the apparent power flows at different lines are inactive.
- The voltage magnitudes are set to their nominal values.

[6] Note that, due to the law of large numbers, the empirical average converges to the expected value as N tends to infinity.

The above assumptions are made to simplify our subsequent presentation. Note that the problem is still highly nonconvex due to the nonconvex power balance equations and the upper and lower bounds on the active power generations. Let $p_{i;k}(\theta) = P_i^g - P_{i;k}^l$ be the net power injection at bus i at time k, where $\theta \in \mathbb{R}^{N-1}$ is a vector collecting the angles at different buses, except for the slack bus. Then the unconstrained objective function can be defined as $\Gamma_k(\theta) = \sum_i^{n_g} c_i(p_{i;k}(\theta) + P_{i;k}^l)$, where $c_i(p_{i;k}(\theta) + P_{i;k}^l)$ is a linear combination of the cost of generation and the penalties on the violation of the lower and upper bound constraints on the active power generation at generator i. Moreover, suppose that $P_{i;k}^l = \bar{P}_i + \gamma_i$, where $\bar{P}$ is a vector collecting the *nominal* loads, and γ is an isotropic random vector with a known distribution $\mathcal{P}$ such that $\mathbb{E}[\gamma_1] = \cdots = \mathbb{E}[\gamma_n] = \bar{\gamma} \neq 0$. In other words, the variations in the load are *biased*. For simplicity of presentation, we abuse the notation and write $\Gamma(\theta; \bar{P} + \gamma^k) = \Gamma_k(\theta)$, where $\gamma^k \sim \mathcal{P}$ is a realization of the randomness in the load at time k. Define the linearization of $\Gamma(\theta; \bar{P} + \gamma)$ around $\bar{P}$ as

$$\Gamma_{\text{lin}}(\theta; \bar{P} + \gamma) = \Gamma(\theta; \bar{P}) + \nabla_P \Gamma(\theta; \bar{P})^\top \gamma \tag{24}$$

For small values of γ, the linearized function $\Gamma_{\text{lin}}(\theta; \bar{P} + \gamma)$ is a good approximation of $\Gamma(\theta; \bar{P} + \gamma)$. In particular, under mild conditions on Γ, the Mean Value theorem implies that $\Gamma(\theta; \bar{P} + \gamma) = \Gamma_{\text{lin}}(\theta; \bar{P} + \gamma) + O(\gamma^2)$. Note that while Γ_{lin} is linear in terms of γ, it is potentially nonconvex with respect to θ. Define effective noise of the linearized functions as

$$\omega_{\text{lin}}^k(\theta; \bar{P}, \gamma^k, \gamma^{k-1}) = \nabla_\theta \Gamma(\theta; \bar{P} + \gamma^k) - \nabla_\theta \Gamma(\theta; \bar{P} + \gamma^{k-1}) \tag{25}$$

for every $k = 1, \ldots, T$. Again, ω_{lin}^k is an accurate approximation of the true effective noise, provided γ is sufficiently small. Note that the bias term in (25) is zero since the right-hand side of (25) has zero mean. Moreover, we can drop the time index k, since the distribution of $\omega_{\text{lin}}^k(\theta; \bar{P}, \gamma^k, \gamma^{k-1})$ does not depend on k, as γ^k and γ^{k-1} are independent and identically distributed. With these definitions, we present our next proposition.

Proposition 1: Suppose that at time T, the objective function of the time-varying OPF corresponds to $\Gamma(\theta; \bar{P})$ with an stationary point $\tilde{\theta}$. Then,

$$\mathbb{E}_{\gamma, \tilde{\gamma} \sim \mathcal{P}}\left[\|\omega_{\text{lin}}(\tilde{\theta}; \bar{P}, \gamma, \tilde{\gamma})\|^2\right] \geq \frac{2\text{Var}_{\gamma \sim \mathcal{P}}(\gamma)}{N} \left(\frac{\left\| \mathbb{E}_{\gamma \sim \mathcal{P}}\left[\nabla_\theta \Gamma_{\text{lin}}(\tilde{\theta}; \bar{P} + \gamma)\right] \right\|}{\bar{\gamma}} \right)^2 \tag{26}$$

A proof of this proposition is in the Appendix. Note that a larger variance of the effective noise leads to a higher smoothing effect, which in turn facilitates the satisfaction of Assumption 2. In essence, Proposition 1 implies that this smoothing effect (captured by the variance of the effective noise) is controlled by the average behavior of the objective function. In particular, suppose that the point $\tilde{\theta}$ is *not* a stationary point of the expected objective function. Therefore, we have $\left\|\mathbb{E}_{\gamma\sim\mathcal{P}}\left[\nabla_\theta\Gamma_{\text{lin}}(\tilde{\theta};\bar{P}+\gamma)\right]\right\| > 0$, and the above proposition implies that the *generalized variance* of the effective noise at $\tilde{\theta}$ increases with the norm of the gradient of the expected function at $\tilde{\theta}$, thereby leading to a higher smoothing effect of the load variation and the elimination of the spurious local minima. This partly explains the high variance of the effective noise at the local minimum of the objective function for the 2-bus system described in Subsection IV-B, and the elimination of its spurious local minimum.

Based on our results, it is possible to eliminate the spurious local solutions in a point-wise OPF problem by adding a synthetically generated noise to the load, thereby elevating the data variation in the problem. This effect of random perturbation in the load values can be observed in Fig. 8c, where it is shown that randomness in the load can eliminate the spurious local minimum and maximum, while keeping the global minimum intact.

However, in practice deriving a class of variation sequences which guarantee convergence is not tractable, due to the nonconvex relationship between the load variation and the "effective noise". This is not surprising, considering the NP-hardness of the time-varying OPF in its worst case. However, even without such a guarantee, computing a discrete OPF trajectory for a load sequence which starts and ends with the load of the target problem may often succeed for a straightforward choice of load variation such as a sinusoidal function, uniform variation, or random walk. Fig. 9 shows two examples on the modified 39-bus system of scaling load by a uniform random walk for 100 time steps with a 20% ramping constraint. As with the sinusoidal load in Fig. 6, we observe that some or all initializations lead to the optimal solution over time, depending on the specific variation.

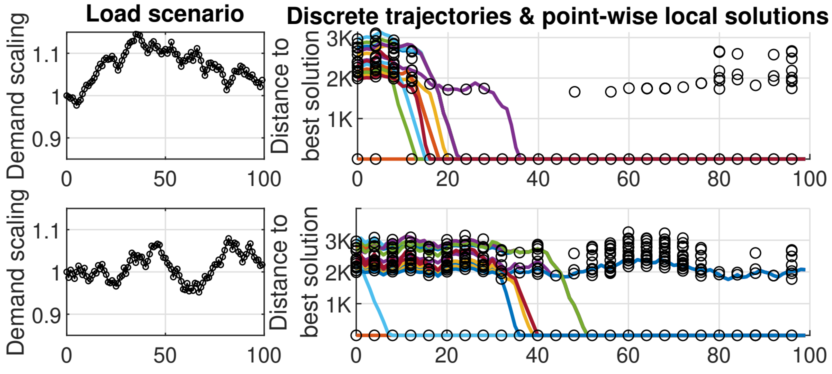


Fig. 9: Numerical studies on the modified 39-bus system with a synthetic load created by a uniform random walk. This type of noise injection could be used to help find the global solution to a static OPF problem.

V. Conclusion

This paper studies time-varying optimal power flow (OPF) problems, in which a set of optimization problems are solved sequentially due to load data variation over time. The solution to each OPF is obtained using local search initialized at the solution to the previous OPF. We offer a case study on a 39-bus system under California data, where the OPF at the initial time has 16 locally optimal solutions leading to 16 solution trajectories. We show that all trajectories converge to the best solution trajectory, even though OPF has many local minima at most times. To understand this highly desirable property, we introduce the notions of escaping period and tracking period, examine the role of data variation and the easiest intermediate problem, study the behavior of the time-varying OPF during the escaping period via a backward-in-time mapping, and relate it to SGD algorithm. By modeling the data variation as a biased noise, we prove that enough data variation enables escaping poor solutions of time-varying OPF over time.

References

[1] J. Mulvaney-Kemp, S. Fattahi, and J. Lavaei, “Load variation enables escaping poor solutions of time-varying optimal power flow,” *IEEE Power & Energy Society General Meeting*, 2020.

[2] J. A. Momoh, *Electric Power System Applications of Optimization*. Boca Raton: CRC Press, /12/19 2017.

[3] R. Baldick, *Applied optimization: formulation and algorithms for engineering systems*. Cambridge University Press, 2006.

[4] W. A. Bukhsh, A. Grothey, K. I. M. McKinnon, and P. A. Trodden, “Local solutions of the optimal power flow problem,” *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 4780–4788, Nov 2013.

[5] R. Y. Zhang, J. Lavaei, and R. Baldick, “Spurious local minima in power system state estimation,” *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 1086–1096, 2019.

[6] R. Y. Zhang, S. Sojoudi, and J. Lavaei, “Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery,” *Journal of Machine Learning research*, 2019.

[7] C. Coffrin, H. L. Hijazi, and P. Van Hentenryck, “The QC relaxation: A theoretical and computational study on optimal power flow,” *IEEE Transactions on Power Systems*, vol. 31, no. 4, pp. 3008–3018, 2015.

[8] B. Kocuk, S. S. Dey, and X. A. Sun, “Strong SOCP relaxations for the optimal power flow problem,” *Operations Research*, vol. 64, no. 6, pp. 1177–1196, 2016.

[9] J. Lavaei and S. H. Low, “Zero duality gap in optimal power flow problem,” *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 92–107, 2012.

[10] S. Sojoudi and J. Lavaei, “Exactness of semidefinite relaxations for nonlinear optimization problems with underlying graph structure,” *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 1746–1778, 2014.

[11] C. Josz, J. Maeght, P. Panciatici, and J. C. Gilbert, “Application of the moment-sos approach to global optimization of the opf problem,” *IEEE Transactions on Power Systems*, vol. 30, no. 1, pp. 463–470, 2014.

[12] D. Kourounis, A. Fuchs, and O. Schenk, “Toward the next generation of multiperiod optimal power flow solvers,” *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4005–4014, 2018.

[13] A. Gopalakrishnan, A. U. Raghunathan, D. Nikovski, and L. T. Biegler, “Global optimization of multi-period optimal power flow,” in *2013 American Control Conference*, June 2013, pp. 1157–1164.

[14] S. Gill, I. Kockar, and G. W. Ault, “Dynamic optimal power flow for active distribution networks,” *IEEE Transactions on Power Systems*, vol. 29, no. 1, pp. 121–131, Jan 2014.

[15] A. Costa and A. S. Costa, “Energy and ancillary service dispatch through dynamic optimal power flow,” *Electric Power Systems Research*, vol. 77, no. 8, pp. 1047 – 1055, 2007.

[16] Y. Tang, K. Dvijotham, and S. Low, “Real-time optimal power flow,” *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2963–2973, 2017.

[17] E. Dall’Anese and A. Simonetto, “Optimal power flow pursuit,” *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 942–952, March 2018.

[18] S. Bolognani, G. Cavraro, and S. Zampieri, *A Distributed Feedback Control Approach to the Optimal Reactive Power Flow Problem*. Heidelberg: Springer International Publishing, 2013, pp. 259–277.

[19] L. Gan and S. H. Low, “An online gradient algorithm for optimal power flow on radial networks,” *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 625–638, March 2016.

[20] S. Fattahi, C. Josz, R. Mohammadi, J. Lavaei, and S. Sojoudi, “Absence of spurious local trajectories in time-varying optimization,” 2019. [Online]. Available: https://lavaei.ieor.berkeley.edu/Time_Varing_2019_1.pdf

[21] M. H. Amini, J. Mohammadi, and S. Kar, “Distributed holistic framework for smart city infrastructures: Tale of interdependent electrified transportation network and power grid,” *IEEE Access*, vol. 7, pp. 157 535–157 554, 2019.

[22] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, “Matpower: Steady-state operations, planning, and analysis tools for power systems research and education,” *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, Feb 2011.

[23] “California ISO OASIS.” [Online]. Available: http://oasis.caiso.com/

[24] O. Massicot and J. Marecek, “On-line non-convex constrained optimization,” *arXiv preprint arXiv:1909.07492*, 2019.

[25] Y. Tang, “Time-varying optimization and its application to power system operation,” Ph.D. dissertation, California Institute of Technology, 2019.

[26] D. Bertsekas, *Nonlinear Programming*, 3rd ed. Belmont, Mass.: Athena Scientific, 2016.

[27] C. B. Do, Q. V. Le, and C. S. Foo, “Proximal regularization for online and batch learning,” *ICML*, 2009.

[28] R. Kleinberg, Y. Li, and Y. Yuan, “An alternative view: When does SGD escape local minima?” *arXiv preprint arXiv:1802.06175*, 2018.

[29] J. Chen and R. Luss, “Stochastic gradient descent with biased but consistent gradient estimators,” *arXiv preprint arXiv:1807.11880*, 2018.

[30] E. K. Ryu and S. Boyd, “Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent,” *http://stanford.edu/~boyd/papers/spi.html*, 2017.

APPENDIX

PROOF OF THEOREM 1

For simplicity of notation, we reverse the order of the time steps, changing $T-t$ to t. Then, the dynamics (13) can be written as

$$\mathbf{z}_t = \mathbf{z}_{t-1} + \eta\nabla\Gamma_0(\mathbf{z}_{t-1}) + \eta\sum_{k=1}^{t-1}\zeta_k(\mathbf{z}_{t-1}) - \eta\sum_{k=1}^{t-1}\omega_k \tag{27}$$

We will extensively use the following sequences of *intermediate* points in our analysis:

$$\mathbf{y}_t = \mathbf{z}_t + \eta\nabla\Gamma_0(\mathbf{z}_t) + \eta\sum_{k=1}^{t}\zeta_k(\mathbf{z}_t) \tag{28}$$

$$\tilde{\mathbf{y}}_t = \mathbf{y}_t - \eta\sum_{k=1}^{t-1}\omega_k \tag{29}$$

It is easy to verify that the above definitions together with (27) gives rise to the following recursive equation:

$$\mathbf{y}_t = \mathbf{y}_{t-1} - \eta\sum_{k=1}^{t-1}\omega_k + \eta\nabla\Gamma_0\left(\mathbf{y}_{t-1} - \eta\sum_{k=1}^{t-1}\omega_k\right) + \eta\sum_{k=1}^{t}\zeta_k\left(\mathbf{y}_{t-1} - \eta\sum_{k=1}^{t-1}\omega_k\right) \tag{30}$$

which in turn implies

$$\mathbf{y}_t = \tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1} + \eta\nabla\Gamma_0\left(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1}\right) + \eta\sum_{k=1}^{t}\zeta_k\left(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1}\right) \tag{31}$$

Define the filtration $\mathcal{F}_{t-1} = \sigma\{\omega_1, \ldots, \omega_{t-2}\}$ and the following stochastic process:

$$G_t = (1+\lambda)^{-t}\left(\|\mathbf{y}_t - \mathbf{z}^*\|^2 - \frac{2(b_1 + b_2 t + b_3 t^2)}{\lambda}\right) \tag{32}$$

where $b_1 := 2\eta^3 r^2 L$, $b_2 := 2\eta^3 r^2 l$, and $b_3 := \frac{4\eta^2 r^2(1+2\lambda)^2}{\lambda}$. Our next lemma provides a lower bound on $\mathbb{E}[\|\mathbf{y}_t - \mathbf{z}^*\|^2 | \mathcal{F}_{t-1}]$ in terms of $\|\mathbf{y}_{t-1} - \mathbf{z}^*\|^2$.

Lemma 1: The following inequality holds:

$$\mathbb{E}[\|\mathbf{y}_t - \mathbf{z}^*\|^2 | \mathcal{F}_{t-1}] \geq (1+\lambda)\|\mathbf{y}_{t-1} - \mathbf{z}^*\|^2 - (b_1 + b_2 t + b_3 t^2) \tag{33}$$

Proof. Based on (31), one can write

$$
\begin{aligned}
&\mathbb{E}[\|\mathbf{y}_t - \mathbf{z}^*\|^2|\mathcal{F}_{t-1}]\\
=&\mathbb{E}[\|\tilde{\mathbf{y}}_{t-1} - \mathbf{z}^* - \eta\omega_{t-1} + \eta\nabla\Gamma_0\left(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1}\right) + \eta\sum_{k=1}^{t}\zeta_k\left(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1}\right)\|^2|\mathcal{F}_{t-1}]\\
\geq&\|\tilde{\mathbf{y}}_{t-1} - \mathbf{z}^*\|^2 + \eta^2\mathbb{E}[\|\omega_{t-1}\|^2|\mathcal{F}_{t-1}] + \mathbb{E}[\|\eta\nabla\Gamma_0(\tilde{\mathbf{y}}_{t-1}-\eta\omega_{t-1}) + \eta\sum_{k=1}^{t}\zeta_k(\tilde{\mathbf{y}}_{t-1}-\eta\omega_{t-1})\|^2|\mathcal{F}_{t-1}]\\
&-2\eta\mathbb{E}[\langle\eta\omega_{t-1}, \nabla\Gamma_0(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1})\rangle|\mathcal{F}_{t-1}]-2\eta\mathbb{E}[\langle\eta\omega_{t-1}, \sum_{k=1}^{t}\Gamma_0(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1})\rangle|\mathcal{F}_{t-1}]\\
&+2\eta\langle\mathbf{z}^* - \tilde{\mathbf{y}}_{t-1}, -\nabla\mathbb{E}[\Gamma_0(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1})|\mathcal{F}_{t-1}]\rangle+2\eta\langle\mathbf{z}^* - \tilde{\mathbf{y}}_{t-1}, -\sum_{k=1}^{t}\mathbb{E}[\zeta_k(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1})|\mathcal{F}_{t-1}]\rangle\\
\geq&\|\tilde{\mathbf{y}}_{t-1} - \mathbf{z}^*\|^2\\
&\underbrace{-2\eta\mathbb{E}[\langle\eta\omega_{t-1}, \nabla\Gamma_0(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1})\rangle|\mathcal{F}_{t-1}]}_{A}\underbrace{-2\eta\mathbb{E}[\langle\eta\omega_{t-1}, \sum_{k=1}^{t}\Gamma_0(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1})\rangle|\mathcal{F}_{t-1}]}_{B}\\
&\underbrace{+2\eta\langle\mathbf{z}^* - \tilde{\mathbf{y}}_{t-1}, -\nabla\mathbb{E}[\Gamma_0(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1})|\mathcal{F}_{t-1}]\rangle}_{C}\underbrace{+2\eta\langle\mathbf{z}^* - \tilde{\mathbf{y}}_{t-1}, -\sum_{k=1}^{t}\mathbb{E}[\zeta_k(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1})|\mathcal{F}_{t-1}]\rangle}_{D}
\end{aligned}
\tag{34}
$$

Next, we will provide a separate lower bound for each term in the above inequality. First, due to Assumption 3, we have

$$C \geq 2\eta c\|\tilde{\mathbf{y}}_{t-1} - \mathbf{z}^*\|^2 \tag{35}$$

and

$$D \geq -2\eta c'\|\tilde{\mathbf{y}}_{t-1} - \mathbf{z}^*\|^2 \tag{36}$$

Furthermore, one can write

$$
\begin{aligned}
A &= -2\eta\mathbb{E}[\langle\eta\omega_{t-1}, \nabla\Gamma_0(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1}) - \nabla\Gamma_0(\tilde{\mathbf{y}}_{t-1})\rangle|\mathcal{F}_{t-1}]\\
&\geq -2\eta\mathbb{E}[\|\eta\omega_{t-1}\|\|\Gamma_0(\tilde{\mathbf{y}}_{t-1} - \eta\omega_{t-1}) - \nabla\Gamma_0(\tilde{\mathbf{y}}_{t-1})\||\mathcal{F}_{t-1}]\\
&\geq -2\eta^3r^2L
\end{aligned}
\tag{37}
$$

where the first equality is due to the fact that $\mathbb{E}[\langle\eta\omega_{t-1}, \Gamma_0(\tilde{\mathbf{y}}_{t-1})\rangle|\mathcal{F}_{t-1}] = 0$. Similarly, we can write

$$B \geq -2\eta^3r^2lt \tag{38}$$

This implies that

$$\begin{aligned}\mathbb{E}[\|\mathbf{y}_t - \mathbf{z}^*\|^2|\mathcal{F}_{t-1}] \geq &(1+2\eta(c-c'))\|\tilde{\mathbf{y}}_{t-1} - \mathbf{z}^*\|^2 - 2\eta^3 r^2(L+lt)\\ =&(1+2\lambda)\|\tilde{\mathbf{y}}_{t-1} - \mathbf{z}^*\|^2 - (b_1+b_2t) \end{aligned} \tag{39}$$

This together with the definition of $\tilde{\mathbf{y}}_{t-1}$ gives rise to the following chain of inequalities

$$\begin{aligned}\mathbb{E}[\|\mathbf{y}_t - \mathbf{z}^*\|^2|\mathcal{F}_{t-1}] \geq &(1+2\lambda)\left\|\mathbf{y}_{t-1} - \eta\sum_{k=1}^{t-2}\omega_k - \mathbf{z}^*\right\|^2 - (b_1+b_2t)\\ \geq &(1+2\lambda)\|\mathbf{y}_{t-1} - \mathbf{z}^*\|^2 - 2(1+2\lambda)\|\mathbf{y}_{t-1} - \mathbf{z}^*\|\left\|\eta\sum_{k=1}^{t-2}\omega_k\right\| - (b_1+b_2t)\\ \geq &(1+2\lambda)\|\mathbf{y}_{t-1} - \mathbf{z}^*\|^2 - 2\eta r(1+2\lambda)t\|\mathbf{y}_{t-1} - \mathbf{z}^*\| - (b_1+b_2t)\end{aligned} \tag{40}$$

Now we consider two cases:

- If $\|\mathbf{y}_{t-1} - \mathbf{z}^*\| \geq \frac{2\eta r(1+2\lambda)t}{\lambda}$, then one can write

$$\mathbb{E}[\|\mathbf{y}_t - \mathbf{z}^*\|^2|\mathcal{F}_{t-1}] \geq (1+\lambda)\|\mathbf{y}_{t-1} - \mathbf{z}^*\|^2 - (b_1+b_2t) \tag{41}$$

- If $\|\mathbf{y}_{t-1} - \mathbf{z}^*\| < \frac{2\eta r(1+2\lambda)t}{\lambda}$, then one can write

$$\mathbb{E}[\|\mathbf{y}_t - \mathbf{z}^*\|^2|\mathcal{F}_{t-1}] \geq (1+2\lambda)\|\mathbf{y}_{t-1} - \mathbf{z}^*\|^2 - \frac{4\eta^2 r^2(1+2\lambda)^2 t^2}{\lambda} - (b_1+b_2t) \tag{42}$$

Combining the above two inequalities leads to

$$\mathbb{E}[\|\mathbf{y}_t - \mathbf{z}^*\|^2|\mathcal{F}_{t-1}] \geq (1+\lambda)\|\mathbf{y}_{t-1} - \mathbf{z}^*\|^2 - (b_1+b_2t+b_3t^2) \tag{43}$$

□

The next lemma is at the crux of our proof for Theorem 1.

Lemma 2: The following two statements hold:

1) G_t is a submartingale with a vanishing drift. More precisely, it satisfies the following inequality

$$\mathbb{E}[G_t|\mathcal{F}_{t-1}] \geq G_{t-1} - (1+\lambda)^{-(t-1)}\left(\frac{2b_2 + 2b_3(2t-1)}{\lambda}\right) \tag{44}$$

2) $\mathbb{E}[G_t] \geq G_0 - \left(\frac{2}{\lambda} + \frac{2}{\lambda^2}\right)b_2 - \left(\frac{4}{\lambda}\left(1+\frac{1}{\lambda}\right)^2\right)b_3$

Proof. One can write

$$\mathbb{E}[G_t|\mathcal{F}_{t-1}] = (1+\lambda)^{-t}\left(\mathbb{E}[\|\mathbf{y}_t - \mathbf{z}^*\|^2|\mathcal{F}_{t-1}] - \frac{2(b_1+b_2t+b_3t^2)}{\lambda}\right) \tag{45}$$

Invoking Lemma 1 leads to

$$\begin{aligned}\mathbb{E}[G_t|\mathcal{F}_{t-1}] \geq &(1+\lambda)^{-t}\Big((1+\lambda)\|\mathbf{y}_{t-1}-\mathbf{z}^*\|^2 - (b_1+b_2t+b_3t^2) - \frac{2(b_1+b_2t+b_3t^2)}{\lambda}\Big)\\ =&(1+\lambda)^{-(t-1)}\|\mathbf{y}_{t-1}-\mathbf{z}^*\|^2 - (1+\lambda)^{-(t-1)}\left(\frac{2(b_1+b_2t+b_3t^2)}{\lambda}\right)\\ =&(1+\lambda)^{-(t-1)}\|\mathbf{y}_{t-1}-\mathbf{z}^*\|^2 - (1+\lambda)^{-(t-1)}\left(\frac{2(b_1+b_2(t-1)+b_3(t-1)^2)}{\lambda}\right)\\ &-(1+\lambda)^{-(t-1)}\left(\frac{2(b_2+b_3(2t-1))}{\lambda}\right)\\ =&G_{t-1}-(1+\lambda)^{-(t-1)}\left(\frac{2b_2+2b_3(2t-1)}{\lambda}\right)\end{aligned} \tag{46}$$

This completes the proof of the first part. To prove the second part, we use the result of the first part together with the tower property of the expectation to write

$$\mathbb{E}[G_t] \geq G_0 - \underbrace{\left(\frac{2b_2}{\lambda}\sum_{k=0}^{t-1}(1+\lambda)^{-k}\right)}_{A} - \underbrace{\left(\frac{4b_3}{\lambda}\sum_{k=0}^{t-1}(k+1)(1+\lambda)^{-k}\right)}_{B} \tag{47}$$

It is easy to verify that

$$A \leq \left(\frac{2}{\lambda}+\frac{2}{\lambda^2}\right)b_2, \qquad B \leq \left(\frac{4}{\lambda}\left(1+\frac{1}{\lambda}\right)^2\right)b_3 \tag{48}$$

This completes the proof. □

Proof of Theorem 1: From the second statement of Lemma 2, one can write

$$\|\mathbf{y}_0-\mathbf{z}^*\|^2 \leq \left(\frac{2}{\lambda}+\frac{2}{\lambda^2}\right)b_2 + \left(\frac{4}{\lambda}\left(1+\frac{1}{\lambda}\right)^2\right)b_3 + (1+\lambda)^{-(t-1)}\,\mathbb{E}[\|\mathbf{y}_{t-1}-\mathbf{z}^*\|^2] \tag{49}$$

On the other hand, one can write

$$\begin{aligned}\mathbb{E}[\|\mathbf{z}_t-\mathbf{z}^*\|^2] &= \mathbb{E}[\|\mathbf{y}_{t-1}-\mathbf{z}^*-\eta\sum_{k=1}^{t-1}\omega_k\|^2]\\ &\geq \mathbb{E}[\|\mathbf{y}_{t-1}-\mathbf{z}^*\|^2] - 2\eta rt\mathbb{E}[\|\mathbf{y}_{t-1}-\mathbf{z}^*\|]\end{aligned} \tag{50}$$

Inequality (50) together with some simple algebra reveals that

$$\mathbb{E}[\|\mathbf{y}_{t-1}-\mathbf{z}^*\|^2] \leq 2\mathbb{E}[\|\mathbf{z}_t-\mathbf{z}^*\|^2] + 16\eta^2r^2t^2 \tag{51}$$

Combining the above inequality with (49) results in

$$\|\mathbf{y}_0-\mathbf{z}^*\|^2 \leq \left(\frac{2}{\lambda}+\frac{2}{\lambda^2}\right)b_2 + \left(\frac{4}{\lambda}\left(1+\frac{1}{\lambda}\right)^2\right)b_3 + 2\,(1+\lambda)^{-(t-1)}\,\mathbb{E}[\|\mathbf{z}_t-\mathbf{z}^*\|^2] + 16\eta^2r^2t^2(1+\lambda)^{-(t-1)} \tag{52}$$

Finally, it only remains to characterize the relationship between $\|\mathbf{y}_0 - \mathbf{z}^*\|^2$ and $\|\mathbf{z}_0 - \mathbf{z}^*\|^2$. To this goal, one can write

$$\begin{aligned}\|\mathbf{y}_0 - \mathbf{z}^*\|^2 &= \|\mathbf{z}_0 - \mathbf{z}^* + \eta\nabla f_0(\mathbf{z}_0)\|^2 \\ &\geq \|\mathbf{z}_0 - \mathbf{z}^*\|^2 - 2\eta\langle \mathbf{z}_0 - \mathbf{z}^*, \eta\nabla f_0(\mathbf{z}_0)\rangle \\ &= \|\mathbf{z}_0 - \mathbf{z}^*\|^2 - 2\eta\langle \mathbf{z}_0 - \mathbf{z}^*, \eta\nabla f_0(\mathbf{z}_0) - \eta\nabla f_0(\mathbf{z}^*)\rangle \\ &\geq \|\mathbf{z}_0 - \mathbf{z}^*\|^2 - 2\eta^2\|\mathbf{z}_0 - \mathbf{z}^*\|\|\nabla f_0(\mathbf{z}_0) - \nabla f_0(\mathbf{z}^*)\| \\ &\geq (1 - 2\eta^2 L)\|\mathbf{z}_0 - \mathbf{z}^*\|^2 \end{aligned} \tag{53}$$

where the last inequality is due to Assumption 1. Combining (53) with (52) concludes the proof. □

Appendix

Proof of Proposition 1

Due to the definition of $\Gamma_{\mathrm{lin}}(\theta; \bar{P} + \gamma)$ in (24), one can write

$$\begin{aligned}\mathbb{E}_{\gamma\sim\mathcal{P}}\left[\nabla_\theta \Gamma_{\mathrm{lin}}(\tilde{\theta}; \bar{P}+\gamma)\right] &= \sum_{i=1}^{N} \nabla_{p_i(\theta)} c_i(p_i(\tilde{\theta})+\bar{P}_i)\nabla_\theta p_i(\tilde{\theta}) + \sum_{i=1}^{N} \nabla_{p_i(\theta)}\nabla_{P_i} c_i(p_i(\tilde{\theta})+\bar{P}_i)\mathbb{E}[\gamma_i]\nabla_\theta p_i(\tilde{\theta}) \\ &= \sum_{i=1}^{N} \nabla_{p_i(\theta)}\nabla_{P_i} c_i(p_i(\tilde{\theta})+\bar{P}_i)\mathbb{E}[\gamma_i]\nabla_\theta p_i(\tilde{\theta}) \end{aligned} \tag{54}$$

where the second equality follows from the assumption $\nabla_\theta \Gamma_{\mathrm{lin}}(\tilde{\theta}; \bar{P}) = 0$. Let us define the vector $v_i = \nabla_{p_i(\theta)}\nabla_{P_i} c_i(p_i(\tilde{\theta})+\bar{P}_i)\nabla_\theta p_i(\tilde{\theta})$. Therefore, one can write

$$\left(\sum_{i=1}^{N} \|v_i\|\right)^2 \geq \left\|\sum_{i=1}^{N} v_i\right\|^2 = \frac{\left\|\mathbb{E}_{\gamma\sim\mathcal{P}}\left[\nabla_\theta \Gamma_{\mathrm{lin}}(\tilde{\theta}; \bar{P}+\gamma)\right]\right\|^2}{\bar{\gamma}^2} \tag{55}$$

On the other hand, a simple calculation reveals that

$$\begin{aligned}\omega^k_{\mathrm{lin}}(\theta; \bar{P}, \gamma, \tilde{\gamma}) &= \nabla_\theta \Gamma(\theta; \bar{P} + \gamma) - \nabla_\theta \Gamma(\theta; \bar{P} + \tilde{\gamma}) \\ &= \sum_{i=1}^{N} \nabla_{p_i(\theta)}\nabla_{P_i} c_i(p_i(\tilde{\theta})+\bar{P}_i)(\gamma_i - \tilde{\gamma}_i)\nabla_\theta p_i(\tilde{\theta}) \end{aligned} \tag{56}$$

Upon defining the matrix $V = [v_1, v_2, \ldots, v_N]$, one can verify that $\omega_{\text{lin}}^k(\theta; \bar{P}, \gamma, \tilde{\gamma}) = V(\gamma - \tilde{\gamma})$, which implies that

$$\begin{aligned}\mathbb{E}_{\gamma,\tilde{\gamma}\sim\mathcal{P}}[\|\omega_{\text{lin}}(\tilde{\theta}; \bar{P}, \gamma, \tilde{\gamma})\|^2] &= \mathbb{E}_{\gamma,\tilde{\gamma}\sim\mathcal{P}}\left[\|V(\gamma - \tilde{\gamma})\|^2\right] \\ &= 2\text{Var}_{\gamma\sim\mathcal{P}}(\gamma)\text{trace}\left(VV^\top\right) \\ &= 2\text{Var}_{\gamma\sim\mathcal{P}}(\gamma)\text{trace}\left(V^\top V\right) \\ &= 2\text{Var}_{\gamma\sim\mathcal{P}}(\gamma)\left(\sum_{i=1}^N \|v_i\|^2\right) \qquad (57)\end{aligned}$$

This implies that

$$\mathbb{E}_{\gamma,\tilde{\gamma}\sim\mathcal{P}}[\|\omega_{\text{lin}}(\tilde{\theta}; \bar{P}, \gamma, \tilde{\gamma})\|^2] \geq \frac{2\text{Var}_{\gamma\sim\mathcal{P}}(\gamma)}{N}\left(\sum_{i=1}^N \|v_i\|\right)^2$$

The above inequality combined with (55) completes the proof. □