Locality in Network Optimization

Patrick Rebeschini  
Yale Institute For Network Science  
Yale University  
New Haven, CT 06511, USA  
patrick.rebeschini@yale.edu

Sekhar Tatikonda  
Yale Institute For Network Science  
Yale University  
New Haven, CT 06511, USA  
sekhar.tatikonda@yale.edu

Abstract

In probability theory and statistics notions of correlation among random variables, decay of correlation, and bias-variance trade-off are fundamental. In this work we introduce analogous notions in optimization, and we show their usefulness in a concrete setting. We propose a general notion of correlation among variables in optimization procedures that is based on the sensitivity of optimal points upon (possibly finite) perturbations. We present a canonical instance in network optimization (the min-cost network flow problem) that exhibits locality, i.e., a setting where the correlation decays as a function of the graph-theoretical distance in the network. In the case of warm-start reoptimization, we develop a general approach to localize a given optimization routine in order to exploit locality. We show that the localization mechanism is responsible for introducing a bias in the original algorithm, and that the bias-variance trade-off that emerges can be exploited to minimize the computational complexity required to reach a prescribed level of error accuracy.

1. Introduction

Many problems in machine learning, networking, control, and statistics can be naturally posed in the framework of network-structured convex optimization. Given the huge problem size involved in modern applications, a lot of efforts have been devoted to the development and analysis of algorithms where the computation and communication are distributed over the network. A crucial challenge remains that of identifying the structural amount of information that each computation node needs to receive from the network to yield an approximate solution with a certain accuracy.

The literature on distributed algorithms in network optimization is gigantic, with much of the earlier seminal work contained in the textbook Bertsekas and Tsitsiklis (1997). More recent work that explicitly relate the convergence behavior of the algorithms being considered to the network topology and size include — but it is certainly not limited to — Johansson et al. (2009); Mosk-Aoyama et al. (2010); Duchi et al. (2012) for first-order methods, and Wei et al. (2013a,b) for second-order methods. Distributed algorithms are iterative in nature. In their synchronous implementations, at every iteration of the algorithm each computation node processes local information that come from its neighbors. Convergence analysis yields the number of iterations that guarantee these algorithms to meet a
prescribed level of error accuracy. In turn, these results translate into bounds on the total amount of information processed by each node, as $k$ iterations of the algorithm means that each node receives information coming from nodes that are at most $k$-hops away in the network. As different algorithms yield different rates of convergence, the bounds on the propagation of information that are so-derived are algorithm-dependent (and also depend on the error analysis being considered). As such, they do not capture structural properties of the optimization problem.

The main aim of this paper is to propose a general notion of “correlation” among variables in network optimization procedures that fundamentally characterizes the propagation of information across the network. This notion is based on the sensitivity of the optimal solution of the optimization problem upon perturbations of parameters locally supported on the network. This notion can be used to investigate “locality,” by which we mean problem instances where the correlation decays as a function of the natural distance in the network, so that, effectively, information only propagates across local portions of it. The phenomenon of locality characterizes situations where local perturbations are confined inside small regions of the network. How small the affected regions are, it depends on the particular problem instance at hand, and on the underlying graph topology. If locality is present, one could hope to exploit it algorithmically. In the case of localized perturbations, for instance, one could hope to localize known optimization routines so that only the affected regions are updated, hence yielding computational savings. While such a localization mechanism would introduce a structural “bias” with respect to the original algorithm — which updates every node in the network and not only the ones that are mostly affected by the perturbation — one could hope to exploit the “reduction in the variance” to achieve a prescribed level of error accuracy with a lower computational complexity. In the following we make all of this precise.

In this paper we define general notions of correlation, decay of correlation (locality), and bias-variance decomposition and trade-off in optimization, and we consider a concrete setting to illustrate their usefulness. The results that we present should be seen as an implementation of a general agenda that can be followed to establish and exploit locality in more general instances of network optimization.

This paper presents four main contributions, discussed in four separate sections, which we now summarize. Proofs are contained in the Appendices.

1) Sensitivity of optimal points. In Section 2 we present a theory on the sensitivity of optimal points for smooth convex problems with linear equality constraints upon (possibly finite) perturbations. We consider the problem of minimizing a smooth convex function $x \rightarrow f(x)$ subject to $Ax = b$, for a certain matrix $A$ and vector $b \in \text{Im}(A)$, where $\text{Im}(A)$ denotes the image of $A$. For the sake of concreteness, we consider this problem as a function of the constraint vector $b$. We prove that if $f$ is strongly convex, then the optimal point $b \rightarrow x^*(b)$ is continuously differentiable along $\text{Im}(A)$, and we explicitly characterize the effect that perturbations have on the optimal solution as a function of the objective function $f$, the constraint matrix $A$ and vector $b$. Given a differentiable function $\varepsilon \in \mathbb{R} \rightarrow b(\varepsilon) \in \text{Im}(A)$, we show that the quantity $\frac{dx^*(b(\varepsilon))}{d\varepsilon}$ is a function of the Moore-Penrose pseudoinverse of the matrix $A\Sigma(b(\varepsilon))A^T$, where $A^T$ is the transpose of $A$, and where $\Sigma(b)$ denotes the inverse of the Hessian of $f$ evaluated at $x^*(b)$, namely, $\Sigma(b) := \nabla^2 f(x^*(b))^{-1}$. The literature on the sensitivity of optimal points (see Section 2 for a list of references related to our setting)
Locality in Network Optimization

is typically only concerned with establishing infinitesimal perturbations locally, i.e., on a neighborhood of a certain \( b \in \text{Im}(A) \). On the other hand, the results that we present extend to finite perturbations as well, as we prove that the derivatives of the optimal point are continuous along \( \text{Im}(A) \), and hence can be integrated to deal with finite perturbations. The workhorse behind our results is Hadamard’s global inverse function theorem (Krantz and Parks 2002), which in our setup yields necessary and sufficient conditions for the inverse of the KKT map to be continuously differentiable. The proofs are given in Appendix A.

2) Notions of correlation in optimization. In Section 3 we provide an interpretation of the sensitivity theory previously developed in terms of notions of correlation among variables in optimization procedures, resembling analogous notions in probability theory. If the matrix \( A \) is full row rank, for instance, then the quantity \( \frac{\partial x^*(b)}{\partial b} \) is well-defined and describes how much \( x^*(b)_i \) — the \( i \)-th component of the optimal solution \( x^*(b) \) — changes upon perturbation of \( b_a \) — the \( a \)-th component of the constraint vector \( b \). We interpret \( \frac{\partial x^*(b)}{\partial b} \) as a measure of the correlation between variables \( i \) and \( a \) in the optimization problem, and we are interested in understanding how this quantity behaves as a function of the geodesic distance between \( i \) and \( a \). We motivate this terminology by establishing an analogy with the theory of correlations in probability, via the connection with Gaussian random variables (the proofs are given in Appendix B). We extend the notion of correlation beyond infinitesimal perturbations, and we show how our theory yields a first instance of comparison theorems for constrained optimization procedures, along the lines of the comparison theorems established in probability theory to capture stochastic decay of correlation and control the difference of high-dimensional distributions (Dobrushin 1970; Rebeschini and van Handel 2014).

In probability theory, decay of correlation characterizes the effective neighborhood dependency of random variables in a probabilistic network. Since the seminal work of Dobrushin (1970), this concept has found many applications beyond statistical physics. Recently, it has been used to develop and prove convergence guarantees for fast distributed local algorithms for inference and decision problems on large networks in a wide variety of domains, for instance, probabilistic inference (Tatikonda and Jordan 2002), wireless communication (Weitz 2006), network learning (Bresler et al. 2008), combinatorial optimization (Gamarnik et al. 2014), and nonlinear filtering (Rebeschini and van Handel 2015). However, even in applications where the underlying problem does not involve randomness, decay of correlation is typically established upon endowing the model with a probabilistic structure, hence modifying the original problem formulation. Our results in network optimization, instead, show how locality can be described in a purely non-random setting, as a structural property of the original optimization problem. To the best of our knowledge, non-random notions of correlation in optimization have been previously considered only in Moallemi and Van Roy (2010), where the authors explicitly use the word “correlation” to denote the sensitivity of optimal points with respect to localized perturbations, and they analyze the correlation as a function of the natural distance in the graph. However, in their work correlation is regarded as a tool to prove convergence guarantees for the specific algorithm at hand (Min-Sum message-passing to solve unconstrained convex problems), and no general theory is built around it.

3) Locality: decay of correlation. As a paradigm for network optimization, in Section 4 we consider the problem of computing network flows. This is a fundamental
problem that has been studied in various formulations by different communities. The min-
cost network flow variant (where the objective function is typically chosen to be linear,
although there are non-linear extensions, and the constraints also include inequalities) has
been essential in the development of the theory of polynomial-times algorithms for opti-
mizations (see Gamarnik et al. (2012) and references therein, or Ahuja et al. (1993) for a
book reference). In the case of quadratic functions, the problem coincides with computing
electrical flows, which is a fundamental primitive related to Laplacian solvers that has been
extensively studied in computer science (see Vishnoi (2013) and references therein) with
also applications to learning problems (Belkin and Niyogi (2004); Kyng et al. (2015), for
instance). In the formulation we consider, a directed graph \( \vec{G} = (V, \vec{E}) \) is given (the bold
notation indicates directed graphs/edges), with its structure encoded in the vertex-edge
incidence matrix \( A \in \mathbb{R}^{V \times \vec{E}} \). To each edge \( e \in \vec{E} \) is associated a flow \( x_e \) with a cost \( f_{e}(x_e) \)
(the function \( f_{e} \) is smooth and strongly convex), and to each vertex \( v \in V \) is associated an
external flow \( b_v \). The network flow problem consists in finding the flow \( x^\star(b) \in \mathbb{R}^{\vec{E}} \) that
minimizes the total cost \( f(x) := \sum_{e \in \vec{E}} f_{e}(x) \) and that satisfies the conservation law
\( Ax = b \).

In this setting, the general sensitivity theory that we have previously developed allows to
calculate the derivatives of the optimal flow in terms of graph Laplacians; in fact, in this
case the matrix \( A \Sigma(b) A^T \) corresponds to the Laplacian of an undirected weighted graph
associated to \( \vec{G} \), where each directed edge \( e \in \vec{E} \) is given a weight \( \Sigma(b)_{ee} > 0 \). Exploiting a
general connection between the Moore-Penrose pseudoinverse of graph Laplacians and the
Green’s function of random walks on weighed graphs — which we present as standalone in
Appendix C — we express the correlation term \( dx^\star(b(\epsilon)) e d\epsilon \) in terms of differences of Green’s
functions. Different graph topologies yield different decaying behaviors for this quantity, as
a function of the geodesic distance \( d(e, Z) \) between edge \( e \) and the set of vertices \( Z \subseteq V \)
where the perturbation is localized, namely, \( Z := \{ z \in V : \frac{db_{z}(\epsilon)}{d\epsilon} = 0 \} \). In the case of ex-
panders, we derive spectral bounds that show an exponential decay, with rate given by the
second largest eigenvalue in magnitude of the diffusion random walk. In this case we estab-
lish various types of decay of correlation bounds, point-to-set and set-to-point. Appendix
D contains the proofs of these results. For grid-like graphs, based on numerical simulations
and on asymptotic results for the behavior of the Green’s function in infinite grids Lawler
and Limic (2010), we expect the correlation term to decay polynomially instead of expo-
nentially.

4) Bias-variance trade-off and scale-free algorithms. In Section 5 we investigate
applications of the framework that we propose to develop scalable computationally-efficient
algorithms. To illustrate the main principle behind our reasoning, we consider the case when
the solution \( x^\star(b) \) is given (also known as warm-start scenario) and we want to compute the
solution \( x^\star(b + p) \) for the perturbed flow \( b + p \), where \( p \) is a perturbation supported on a
small localized subset \( Z \subseteq V \). This setting belongs to the class of reoptimization problems
typically studied in computer science and operations research (however, we were not able to
find previous results on the particular instance that we consider; the cases considered
more frequently involve discrete problems and changes in the graph structure, in the spirit
of Archetti et al. (2003), for instance).

The decay of correlation property structurally exhibited by the min-cost network flow prob-
lem encodes the fact that when the external flow \( b \) is locally perturbed it suffices to re-
compute the solution only for the part of the network that is “mostly affected” by this perturbation, i.e., the set of edges that have a distance at most $r$ from the perturbation set $Z$. Here, the radius $r$ is tuned to meet the desired level of error tolerance given the size of the perturbation and given the graph topology being investigated. This allows us to show that it is possible to develop localized versions of canonical optimization algorithms that can exploit locality by only updating the edges in a subgraph of $\vec{G}$. For the sake of concreteness, we investigate the convergence properties of a localized version of the projected gradient descent algorithm, a canonical routine in optimization. Analogous results can be established for any optimization routine. We show that localized algorithms can yield scale-free results, in the sense that the computational complexity required to reach a prescribed level of error accuracy does not depend on the dimension of the whole network $\vec{G}$. This is in sharp contrast to the complexity of “global” algorithms that do not exploit locality and update the solution at every edge in $\vec{G}$.

The key to establish scale-free results is a bias-variance decomposition that we give for the error of localized algorithms. This decomposition exemplifies the idea that by introducing some bias in a given optimization routine (in our setting, by truncating the problem size restricting the algorithm to a subgraph) one can diminish its variance and obtain faster convergence rates. Proofs of the results here presented are contained in Appendix E.

The theory that we develop in the simple context of smooth strongly convex optimization problems with linear constraints and applications to network flows illustrates a general framework to study the trade-off between accuracy and complexity in local algorithms for network optimization.

**Remark 1 (Notation)** Throughout, for a given real matrix $M$, we denote by $M^T$ its transpose, by $M^{-1}$ its inverse, and by $M^+$ its Moore-Penrose pseudoinverse. We denote by $\text{Ker}(M) := \{x : Mx = 0\}$ and $\text{Im}(M) := \{y : y = Mx \text{ for some } x\}$ the kernel and the image of $M$, respectively. Given an index set $\mathcal{I}$ and subsets $K,L \subseteq \mathcal{I}$, if $M \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$, we let $M_{K,L} \in \mathbb{R}^{K \times L}$ denote the submatrix corresponding to the rows of $M$ indexed by $K$ and the columns of $M$ indexed by $L$. We use the notation $I$ to indicate the identity matrix, $1$ to indicate the all-one vector (or matrix), and $0$ to indicate the all-zero vector (or matrix), whose sizes will be implied by the context. Given a vector $x \in \mathbb{R}^I$, we denote by $x_i \in \mathbb{R}$ the component associated to $i \in \mathcal{I}$, and by $x_K := (x_i)_{i \in K} \in \mathbb{R}^K$ the components associated to a subset $K \subseteq \mathcal{I}$. Let $\|x\| := (\sum_{i \in \mathcal{I}} x_i^2)^{1/2}$ denote the $\ell_2$-norm of $x$, and define the localized $\ell_2$-norm on $K$ as $\|x\|_K := (\sum_{i \in K} x_i^2)^{1/2}$. Clearly, $\|x\|_K = \|x_K\|$ and $\|x\|_I = \|x\|$. We use the notation $|K|$ to denote the cardinality of $K$. If $\vec{G} = (V, \vec{E})$ denotes a directed graph with vertex set $V$ and edge set $\vec{E}$, we let $G = (V,E)$ represent the undirected graph naturally associated to $\vec{G}$, namely, $\{u,v\} \in E$ if and only if either $(u,v) \in \vec{E}$ or $(v,u) \in \vec{E}$.

**2. Sensitivity of optimal points**

Let $\mathcal{V}$ be a finite set — to be referred to as the “variable set” — and let $f : \mathbb{R}^\mathcal{V} \to \mathbb{R}$ be a strictly convex function, twice continuously differentiable. Let $\mathcal{F}$ be a finite set — to be referred to as the “factor set” — and let $A \in \mathbb{R}^{\mathcal{F} \times \mathcal{V}}$. Consider the following optimization
problem over $x \in \mathbb{R}^V$:

$$
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad Ax = b,
\end{align*}
$$

for $b \in \text{Im}(A) \subseteq \mathbb{R}^F$, so that the feasible region is not empty. Throughout this paper we think of the function $f$ and the matrix $A$ as fixed, and we consider the solution of the optimization problem above as a function of the vector $b \in \text{Im}(A)$. By strict convexity, this problem has a unique optimal solution, that we denote by $x^*(b) := \arg \min \{f(x) : x \in \mathbb{R}^V, Ax = b\}$.

Theorem 2 below provides a characterization of the way a perturbation of the constraint vector $b$ along the subspace $\text{Im}(A)$ affects the optimal solution $x^*(b)$ in the case when the function $f$ is strongly convex. The proof is given in Appendix A.

**Theorem 2 (Sensitivity optimal point)** Let $f : \mathbb{R}^V \to \mathbb{R}$ be a strongly convex function, twice continuously differentiable. Let $A \in \mathbb{R}^{F \times V}$. For $b \in \text{Im}(A)$, let $\Sigma(b) := \nabla^2 f(x^*(b))^{-1}$ and $D(b) := \Sigma(b)A^T(A\Sigma(b)A^T)^+$. Then, $x^*$ is continuously differentiable along the subspace $\text{Im}(A)$, and given a differentiable function $\varepsilon \in \mathbb{R} \to b(\varepsilon) \in \text{Im}(A)$ we have

$$
\frac{dx^*(b(\varepsilon))}{d\varepsilon} = D(b(\varepsilon)) \frac{db(\varepsilon)}{d\varepsilon}.
$$

Most of the literature on the sensitivity of optimal points for nonlinear programs investigates **local** results for infinitesimal perturbations, establishing the existence of a local neighborhood of the perturbed parameter(s) where the optimal point(s) has(have) some analytical properties such as continuity, differentiability, etc. Classical references are Guddat (1976); Robinson (1979); Fiacco (1983); Kyparisis (1986). Typically, the main tool used to establish this type of results is the implicit function theorem applied to the first order optimality conditions, which is a local statement (as such, it is flexible and it can be applied to a great variety of optimization problems). The sensitivity result that we present in Theorem 2 in the case of strongly convex functions, on the other hand, is based on Hadamard’s global inverse function theorem Krantz and Parks (2002), and it is a **global** statement, as it shows that the optimal point $x^*$ is continuously differentiable along the entire subspace $\text{Im}(A)$. This fact allows the use of the fundamental theorem of calculus to deal with finite perturbations, which is instrumental for the results developed in this paper (in particular, for the connection with comparison theorems in probability, Section 3.2, and for the results in Section 5). In the case of quadratic programs with linear constraints (the problem we consider can be thought of as an extension of this setting to strongly convex functions) there are many papers in parametric programming investigating the sensitivity of optimal points with respect to changes in the objective functions and constraints (see Phu and Yen (2001) and references therein). While most of these results are again local, some of them are closer to our approach and also address finite perturbations Boot (1963).

Theorem 2 characterizes the behavior of $x^*(b)$ upon perturbations of $b$ along $\text{Im}(A) \subseteq \mathbb{R}^F$. If the matrix $A$ is full row rank, i.e., $\text{Im}(A) = \mathbb{R}^F$, then the optimal point $x^*$ is everywhere continuously differentiable, and we can compute its gradient. We have the following immediate corollary.
Corollary 3 (Sensitivity of the optimal point, full rank case) Consider the setting of Theorem 2, with the matrix $A \in \mathbb{R}^{F \times V}$ having full row rank, i.e., $\text{Im}(A) = \mathbb{R}^F$. Then, the function $b \in \mathbb{R}^F \to x^*(b) \in \mathbb{R}^V$ is continuously differentiable and

$$\frac{dx^*(b)}{db} = D(b) = \Sigma(b)A^T(A\Sigma(b)A^T)^{-1}.$$

3. Notions of correlation in optimization

The sensitivity analysis presented in Section 2 suggests a natural notion of correlation between variables and factors in optimization, resembling notions of correlation among random variables in probability theory. If the matrix $A$ is full row rank, then the quantity $\frac{\partial x^*(b)}{\partial b}$ is well-defined and it captures the interaction between variable $i \in V$ and factor $a \in F$ in the optimization procedure, and the quantity $D(b)_{ia}$ in Corollary 3 characterizes this correlation as a function of the constraint matrix $A$, the objective function $f$, and the optimal solution $x^*(b)$. Theorem 2 allows us to extend the notion of correlation between variables and factors to the more general case when the matrix $A$ is not full rank. As an example, let $b, p \in \text{Im}(A)$, and assume that $p$ is supported on a subset $Z \subseteq F$, namely, $p_a \neq 0$ if and only if $a \in Z$. Define $b(\varepsilon) := b + \varepsilon p$. Then, the quantity $\frac{dx^*(b(\varepsilon))}{d\varepsilon}$ measures how much a perturbation of the constraints in $Z$ affects the optimal solution at $i \in V$, hence it can be interpreted as a measure of the correlation between variable $i$ and the factors in $Z$, which is characterized by the term $(D(b(\varepsilon))_{ia})_{i} = \sum_{a \in Z} D(b(\varepsilon))_{ia}p_a$ in Theorem 2.

We now discuss how these notions of correlation relate to analogous notions in probability.

3.1 Connection with Gaussian random variables

The resemblance between the notion of correlations in optimization and in probability is made explicit via the analogy to the theory of Gaussian random variables. Recall the following result (proofs of the results here discussed are given for completeness in Appendix B).

Proposition 4 (Conditional mean of Gaussian random variables) Let $V, F$ be two finite sets. Let $X \in \mathbb{R}^V$ be a Gaussian random vector with mean $\mu \in \mathbb{R}^V$ and covariance $\Sigma \in \mathbb{R}^{V \times V}$, possibly singular. Let $A \in \mathbb{R}^{F \times V}$ be given. Given a differentiable function $\varepsilon \in \mathbb{R} \to b(\varepsilon) \in \text{Im}(A)$, we have $\frac{d\mathbb{E}[X|AX=b(\varepsilon)]}{d\varepsilon} = \Sigma A^T(\Sigma A^T)^{-1} + \frac{db(\varepsilon)}{d\varepsilon}$. If $\Sigma$ is invertible and $A$ is full row rank, then for each $b \in \mathbb{R}^F$ we have $\frac{d\mathbb{E}[X|AX=b]}{db} = \Sigma A^T(\Sigma A^T)^{-1}$.

Proposition 4 shows that for $i \in V, a \in F$, the quantity $\frac{\partial \mathbb{E}[X_i|AX=b]}{\partial b}$ can be interpreted as a measure of correlation between the random variables $X_i$ and $(AX)_a$, as it describes how much a perturbation of $(AX)_a$ impacts $X_i$, upon conditioning on $AX$. A similar interpretation can be given in optimization for the quantities in Corollary 3, with the difference that typically these quantities depend on $b \in \text{Im}(A)$, as they are functions of $x^*(b)$.

The sensitivity results that we derived in Section 2 also yield notions of correlation in optimization between variables. The following lemma, an immediate application of Corollary 3, shows that the local behavior of the optimal solution of the optimization problem
(1) when we freeze some coordinates, upon perturbation of these coordinates, is analogous to the behavior of the conditional mean of a non-degenerate Gaussian random vector upon changing the coordinates we condition on. Recall (see the proof of Proposition 4) that if $X \in \mathbb{R}^V$ is a Gaussian vector with mean $\mu \in \mathbb{R}^V$ and positive definite covariance $\Sigma \in \mathbb{R}^{V \times V}$, then for $I \subseteq V$ and $B := V \setminus I$, \[
abla^2 \mathbb{E}[X_I | X_B = x_B] = \Sigma_{I,B}(\Sigma_{B,B})^{-1}.\]

**Lemma 5 (Sensitivity with respect to boundary conditions)** Let $f : \mathbb{R}^V \rightarrow \mathbb{R}$ be a strongly convex function, twice continuously differentiable. Let $I \subseteq V$ be a not empty set, and let $B := V \setminus I$ not empty. Define the function $x_I^* : x_B \in \mathbb{R}^B \rightarrow x_I^*(x_B) := \arg \min \{ f(x_I x_B) : x_I \in \mathbb{R}^I \}$. For $x_B \in \mathbb{R}^B$, let $H(x_B) := \nabla^2 f(x_I^*(x_B) x_B)$ and $\Sigma(x_B) := H(x_B)^{-1}$. Then, $x_I^*$ is continuously differentiable and \[
abla x_I^*(x_B) = \Sigma(x_B)_{I,B}(\Sigma(x_B)_{B,B})^{-1} = -(H(x_B)_{I,I})^{-1}H(x_B)_{I,B}.\]

### 3.2 Comparison theorems

The connection between Theorem 2 and the theory of correlations in probability extends beyond infinitesimal perturbations. As previously discussed, Theorem 2 can be used to deal with finite perturbations, and so it can be interpreted as a comparison theorem to capture uniform correlations in optimization, along the lines of the comparison theorems in probability theory to capture stochastic decay of correlation and control the difference of high-dimensional distributions (see the seminal work in Dobrušin (1970), and Rebeschini and van Handel (2014) for generalizations).

To see this analogy, let us consider a simplified version of the Dobrushin comparison theorem that can be easily derived from the textbook version in Georgii (2011), Theorem 8.20. Let $I$ be a finite set, and let $\Omega := \prod_{i \in I} \Omega_i$ where $\Omega_i$ is a finite set for each $i \in I$. Define the projections $X_i : x \mapsto x_i$ for $x \in \Omega$ an $i \in I$. For any probability distribution $\mu$ on $\Omega$, define the marginal $\mu_i(y) := \mu(X_i = y)$, and the conditional distribution $\mu_{ij}(y) := \mu(X_i = y | X_{I\setminus\{i\}} = x_{I\setminus\{i\}})$. Define the total variation distance between two distributions $\nu$ and $\tilde{\nu}$ on $\Omega_i$ as $\|\nu - \tilde{\nu}\|_T := \frac{1}{2} \sum_{y \in \Omega_i} |\nu(y) - \tilde{\nu}(y)|$.

**Theorem 6 (Dobrushin comparison theorem)** Let $\mu, \tilde{\mu}$ be probability distributions on $\Omega$. For each $i,j \in I$, define $C_{ij} := \sup_{x,z \in \Omega_{I\setminus\{i\}} \setminus \{y\}} \|\mu_{ij} - \mu_{ij}^z\|_T$ and $b_j := \sup_{x \in \Omega_i} \|\mu_j^x - \tilde{\mu}_j\|_T$, and assume that the Dobrushin condition holds: $\max_{i,j} \sum_{j \in I} C_{ij} < 1$. Then the matrix sum $D := \sum_{t \geq 0} C^t$ is convergent, and for any $i \in I$, $y \in \Omega_i$, we have $\|\mu_i - \tilde{\mu}_i\|_T \leq \sum_{j \in I} D_{ij} b_j$.

The Dobrushin coefficient $C_{ij}$ is a (uniform) measure of the degree to which a perturbation of site $j$ directly affects site $i$ under the distribution $\mu$. However, perturbing site $j$ might also indirectly affect site $i$: it could affect another site $k$ which in turn affects $i$, etc. The aggregate effect of a perturbation at site $j$ on site $i$ is captured by $D_{ij}$. The quantity $b_j$ is a comparison term that measures the local difference at site $j$ between $\mu$ and $\tilde{\mu}$ (in terms of the conditional distributions $\mu_j^x$ and $\tilde{\mu}_j$).

The formal analogy between the Dobrushin comparison theorem and the sensitivity results of Theorem 2 for the optimization problem (1) is made explicit by the fundamental theorem of calculus. This connection is easier to make if we assume that the matrix $A$ has full row rank, and we consider the results in Corollary 3. In this setting, the optimal
point \( x^* \) is everywhere continuously differentiable, and for each \( i \in \mathcal{V}, b, \tilde{b} \in \mathbb{R}^F, b \neq \tilde{b}, \) we have \( x^*(b)_i - x^*(\tilde{b})_i = \int_0^1 dx^*(\theta b + (1-\theta)\tilde{b})_i = \sum_{a \in \mathcal{F}} D(b, \tilde{b})_{ia} (b_a - \tilde{b}_a), \) where \( D(b, \tilde{b})_{ia} := \int_0^1 (\Sigma(b_\theta)A^T(\Sigma(b_\theta)A^T)^{-1})_{ia} d\theta \) with \( b_\theta := \theta b + (1-\theta)\tilde{b}. \) If for each \( i \in \mathcal{V} \) and \( a \in \mathcal{F} \) we have \( \sup_{b \in \mathbb{R}^F} |(\Sigma(b)A^T(\Sigma(b)A^T)^{-1})_{ia}| \leq D_{ia}, \) then from the previous expression we find \( |x^*(b)_i - x^*(\tilde{b})_i| \leq \sum_{a \in \mathcal{F}} D_{ia} |b_a - \tilde{b}_a|, \) whose structure resembles the statement in Theorem 6. The quantity \( D_{ia} \) represents a uniform measure of the aggregate impact that a perturbation of the \( a \)-th component of the constraint vector \( b \) has to the \( i \)-th component of the optimal solution \( x^* \), so the matrix \( D \) takes the analogous role of the matrix \( \mathcal{D} \) in Theorem 6 (and as we will see below in a concrete application, see Theorem 7, suitable series expansions of \( D \) yield the analogous of \( \mathcal{C} \)). The quantity \( |b_a - \tilde{b}_a| \) is a comparison term that measures the local difference at factor \( a \) between \( b \) and \( \tilde{b} \), resembling the role of \( \delta_j \) in Theorem 6.

In the next section we investigate the notion of correlation just introduced in the context of network optimization, in a concrete instance when the constraints naturally reflect a graph structure, and we investigate the behavior of the correlations as a function of the natural distance in the graph.

### 4. Locality: decay of correlation

As a paradigm for network optimization, we consider the network flow problem that has been widely studied in various fields (see introduction). Consider a directed graph \( \tilde{G} := (\mathcal{V}, \tilde{E}) \), with vertex set \( \mathcal{V} \) and edge set \( \tilde{E} \), with no self-edges and no multiple edges. Let \( G = (\mathcal{V}, E) \) be the undirected graph naturally associated with \( \tilde{G} \), that is, \( \{u, v\} \in E \) if and only if either \( \{u, v\} \in \tilde{E} \) or \( \{v, u\} \in \tilde{E} \). Without loss of generality, assume that \( G \) is connected (otherwise we can treat each connected component on its own). For each \( e \in \tilde{E} \) let \( x_e \) denote the flow on edge \( e \), with \( x_e > 0 \) if the flow is in the direction of the edge, \( x_e < 0 \) if the flow is in the direction opposite the edge. For each \( v \in \mathcal{V} \) let \( b_v \) be a given external flow on the vertex \( v: b_v > 0 \) represents a source where the flow enters the vertex, whereas \( b_v < 0 \) represents a sink where the flow enters the vertex. Assume that the total of the source flows equals the total of the sink flows, that is, \( \mathbf{1}^T b = \sum_{v \in \mathcal{V}} b_v = 0 \), where \( b = (b_v)_{v \in \mathcal{V}} \in \mathbb{R}^\mathcal{V} \) is the flow vector. We assume that the flow satisfies a conservation equation so that at each vertex the total flow is zero. This conservation law can be expressed as \( Ax = b \), where \( A \in \mathbb{R}^{\mathcal{V}\times\tilde{E}} \) is the vertex-edge incidence matrix defined as \( A_{ve} := 1 \) if \( e \) leaves node \( v \), \( A_{ve} := -1 \) if \( e \) enters node \( v \), and \( A_{ve} := 0 \) otherwise. For each edge \( e \in \tilde{E} \) let \( f_e : \mathbb{R} \to \mathbb{R} \) be its associated cost function, assumed to be strongly convex and twice continuously differentiable. The network flow problem reads as problem (1) with \( f(x) := \sum_{e \in \tilde{E}} f_e(x_e) \). It is easy to see that since \( G \) is connected \( \text{Im}(A) \) consists of all vectors orthogonal to the vector \( \mathbf{1} \), i.e., \( \text{Im}(A) = \{y \in \mathbb{R}^\mathcal{V} : \mathbf{1}^Ty = 0\} \). Henceforth, for each \( b \in \mathbb{R}^\mathcal{V} \) with \( \mathbf{1}^T b = 0 \), let \( x^*(b) \) be the optimal flow.

We first apply the sensitivity theory developed in Section 2 to characterize the correlation between vertices (i.e., factors) and edges (i.e., variables) in the network flow problem. Then, we investigate the behavior of these correlations in terms of the natural distance on \( G \).
4.1 Correlation, graph Laplacians and Green’s functions

In the setting of the network flow problem, Theorem 2 immediately allows us to characterize the derivatives of the optimal point $x^*$ along the subspace $\text{Im}(A)$ as a function of the graph Laplacian Chung (1997). For $b \in \mathbb{R}^V$ such that $1^T b = 0$, let $\Sigma(b) := \nabla^2 f(x^*(b))^{-1} \in \mathbb{R}^E \times \overline{E}$, which is a diagonal matrix with entries given by $\sigma(b)_e := \sum(b)_{ee} := (\frac{\partial^2 f_e(x^*(b))}{\partial x^2})^{-1} > 0$. Each term $\sigma(b)_e$ is strictly positive as $f_e$ is strongly convex by assumption. Let $W(b) \in \mathbb{R}^{V \times V}$ be the symmetric matrix defined, for each $u, v \in V$, as $W(b)_{uv} := \sigma(b)_e$ if $e = (u, v) \in \overline{E}$ or $e = (v, u) \in \overline{E}$, and $W(b)_{uv} := 0$ otherwise. Let $D(b) \in \mathbb{R}^{V \times V}$ be the diagonal matrix with entries given by $d(b)_v := D(b)_{vv} := \sum_{u \in V} W(b)_{vu}$, for $v \in V$. Let $L(b) := D(b) - W(b)$ be the graph Laplacian of the undirected weighted graph $(V, E, W(b))$, where to each edge $e = \{u, v\} \in E$ is associated the weight $W(b)_{uv}$. A direct application of Theorem 2 (upon choosing variable set $V := \overline{E}$ and factor set $F := V$, and noticing that $A \Sigma(b) A^T = L(b)$) shows that the derivatives of the optimal point $x^*$ along $\text{Im}(A)$ can be expressed in terms of the Moore-Penrose pseudoinverse of $L(b)$. The connection between $L(b)^+$ and the Green’s function of random walks with transition matrix $P(b) := (D(b))^{-1} W(b)$ allows us to derive the following result (proofs are in Appendix C).

**Theorem 7 (Sensitivity optimal flow)** For $b \in \mathbb{R}^V$, $1^T b = 0$, let $D(b) := \Sigma(b) A^T L(b)^+$. The optimal network flow $x^*$ is continuously differentiable along $\text{Im}(A)$, and given a differentiable function $\varepsilon \in \mathbb{R} \to b(\varepsilon) \in \text{Im}(A)$ we have $\frac{dx^*(b(\varepsilon))}{d\varepsilon} = D(b(\varepsilon)) \frac{db(\varepsilon)}{d\varepsilon}$. For $e = \{u, v\} \in \overline{E}$,

$$\frac{dx^*(b(\varepsilon))}{d\varepsilon} = \sigma(b)_e \sum_{z \in V} \frac{1}{d(b)_z} \frac{db(\varepsilon)_z}{d\varepsilon} \sum_{t=0}^{\infty} (P(b)_u)^t \frac{db(\varepsilon)_z}{d\varepsilon} (P(b)_v)^t - (P(b)_v)^t (P(b)_u)^t).$$

Let $b, p \in \mathbb{R}^V$ such that $1^T b = 1^T p = 0$, and assume that $p$ is supported on a subset $Z \subseteq V$, namely, $p_v \neq 0$ if and only if $v \in Z$. Define $b(\varepsilon) := b + \varepsilon p$. Then, as discussed in Section 3, the quantity $\frac{dx^*(b(\varepsilon))}{d\varepsilon}$ can be interpreted as a measure of the correlation between edge $e \in \overline{E}$ and the vertices in $Z$ in the network flow problem. How does the correlation behave with respect to the graph distance between $e$ and $Z$? Theorem 7 shows that this behavior is controlled by the difference of the Green’s function $\sum_{t=0}^{\infty} P(b)^t_{uz}$ with respect to two neighbor starting points $u$ and $v$ (note that the Green’s function itself is infinite, as we are dealing with finite graphs). Different graph topologies yield different decaying behaviors for this quantity. In the case of expanders, we now derive spectral bounds that decay exponentially, with rate given by the second largest eigenvalue in magnitude of the diffusion random walk. For grid-like topologies, based on simulations and on asymptotic results for the behavior of the Green’s function in infinite grids Lawler and Limic (2010), we expect the correlation to decay polynomially instead of exponentially.

4.2 Decay of correlation for expanders

Let $n := |V|$ be the cardinality of $V$, and for each $b \in \text{Im}(A)$ let $-1 \leq \lambda_n(b) \leq \lambda_{n-1}(b) \leq \cdots \leq \lambda_2(b) < \lambda_1(b) = 1$ be the real eigenvalues of $P(b)$.\footnote{This characterization of eigenvalues for random walks on connected weighted graphs follows from the Perron-Frobenius theory. See Lovász (1993).} Define $\lambda(b) := \max\{|\lambda_2(b)|, |\lambda_n(b)|\}$.
and $\lambda := \sup_{b \in \text{Im}(A)} \lambda(b)$. For each $v \in V$, let $\mathcal{N}(v) := \{w \in V : \{v, w\} \in E\}$ be the set of node neighbors of $v$ in the graph $G$. Let $d$ be the graph-theoretical distance between vertices in the graph $G$, namely, $d(u, v)$ is the length of the shortest path between vertices $u, v \in V$. For subset of vertices $U, Z \subseteq V$, define $d(U, Z) := \min\{d(u, z) : u \in U, z \in Z\}$. For each subset of edges $\vec{F} \subseteq \vec{E}$, let $V_{\vec{F}} \subseteq V$ be the vertex set of the subgraph $(V_{\vec{F}}, \vec{F})$ of $\vec{G}$ that is induced by the edges in $\vec{F}$. Recall the definition of the localized $\ell_2$-norm from Remark 1. The following result attests that the correlation for the network flow problem is upper-bounded by a quantity that decays exponentially as a function of the distance in the graph, with rate given by $\lambda$. For graphs where $\lambda$ does not depend on the dimension, i.e., expanders, Theorem 8 can be interpreted as a first manifestation of the decay of correlation principle (i.e., locality) in network optimization. The proof is given in Appendix D.

**Theorem 8 (Decay of correlation for expanders)** Consider the setting defined above. Let $\varepsilon \in \mathbb{R} \rightarrow b(\varepsilon) \in \text{Im}(A)$ be a differentiable function such that for any $\varepsilon \in \mathbb{R}$ we have $\frac{db(\varepsilon)}{d\varepsilon} \neq 0$ if and only if $v \in Z$, for a given $Z \subseteq V$. Then, for any subset of edges $\vec{F} \subseteq \vec{E}$ and any $\varepsilon \in \mathbb{R}$, we have

$$\left\| \frac{d \ast (b(\varepsilon))}{d\varepsilon} \right\|_{\vec{F}} \leq c \frac{\lambda^{d(V_{\vec{F}}, Z)}}{1 - \lambda} \left\| \frac{db(\varepsilon)}{d\varepsilon} \right\|_{Z},$$

with $c := \sup_{b \in \text{Im}(A)} \frac{\max_{v \in V_{\vec{F}}} \sqrt{2|\mathcal{N}(v) \cap V_{\vec{F}}|}}{\min_{v \in V_{\vec{F}}} d(b)_v} \max_{u, v \in V_{\vec{F}}} W(b)_{uv}$.

Recall that $\left\| \frac{d \ast (b(\varepsilon))}{d\varepsilon} \right\|_{\vec{F}} \equiv (\sum_{e \in \vec{F}} \left(\frac{d \ast (b(\varepsilon))}{d\varepsilon}\right)_e)^{1/2}$ and $\left\| \frac{db(\varepsilon)}{d\varepsilon} \right\|_{Z} \equiv (\sum_{v \in Z} \left(\frac{db(\varepsilon)}{d\varepsilon}\right)_v)^{1/2}$. The bound in Theorem 8 controls the effect that a localized perturbation supported on a subset of vertices $Z \subseteq V$ has on a subset of edges $\vec{F} \subseteq \vec{E}$, as a function of the distance between $\vec{F}$ and $Z$, i.e., $d(V_{\vec{F}}, Z)$ (note that we only defined the distance among vertices, not edges, and that this distance is with respect to the unweighted graph $G$). See Figure 1. A key feature of Theorem 8 — which is essential for the results in Section 5 below — is that the bound presented does not depend on the cardinality of $\vec{F}$.

Figure 1: Representation of a directed graph $\vec{G} = (V, \vec{E})$, a subset of edges $\vec{F} \subseteq \vec{E}$ (edges colored in green), and a subset of vertices $Z \subseteq V$ (vertices colored in red). The vertices in $V_{\vec{F}}$, the vertex set induced by the edge set $\vec{F}$, are colored in green. Here, $d(V_{\vec{F}}, Z) = 3$. 

11
Theorem 8 controls the effect that a single localized perturbation (supported on multiple vertices, as it has to be that \(|Z| \geq 2\) for the function \(\varepsilon \in \mathbb{R} \to b(\varepsilon)\) to be on \(\text{Im}(A)\)) has on a collection of edges for the optimal solution, independently of the number of edges being considered. We refer to this type of decay of correlation as set-to-point. Analogously, it is possible to control the effect that multiple localized perturbations have on a single edge for the optimal solution, independently of the number of perturbations being considered. We refer to this type of decay of correlation as point-to-set. To illustrate in more detail these two types of decay of correlation, and for the sake of simplicity, we consider perturbations that are supported on exactly two vertices, corresponding to the endpoints of edges. Given \(b \in \mathbb{R}^V\) such that \(1^Tb = 0\), and \(e = (u, v) \in \vec{E}\), we define the directional derivative of \(x^*\) along edge \(e\) evaluated at \(b\) as

\[
\nabla_{e}x^*(b) := \frac{dx^*(b+\varepsilon e_u-e_v))}{d\varepsilon}|_{\varepsilon=0},
\]

where for each \(v \in V\), \(e_v \in \mathbb{R}^V\) is the vector defined as \((e_v)_w = 0\) if \(w \neq v\) and \((e_v)_v = 1\). Then, we immediately have the following corollary of Theorem 8.

**Corollary 9 (Set-to-point decay of correlation)** For \(b \in \mathbb{R}^V\) with \(1^Tb = 0\), \(\vec{F} \subseteq \vec{E}\) and \(e \in \vec{E}\), we have

\[
\|\nabla_{e}x^*(b)\|_{\vec{F}} \equiv \sqrt{\sum_{f \in \vec{F}}(\nabla_{e}x^*(b)_f)^2} \leq \sqrt{2c} \frac{\lambda^d(V_{\vec{F}},V_{e})}{1-\lambda},
\]

where \(c\) is as in Theorem 8.

The key feature of the bound in Corollary 9 is that it does not depend on the cardinality of \(\vec{F}\). Exploiting the symmetry of the identities involving the graph Laplacian, it is also easy to establish the following analogous result. The proof is given in Appendix D.

**Lemma 10 (Point-to-set decay of correlation)** For \(b \in \mathbb{R}^V\) with \(1^Tb = 0\), \(\vec{F} \subseteq \vec{E}\) and \(f \in \vec{E}\), we have

\[
\sqrt{\sum_{e \in \vec{F}}(\nabla_{e}x^*(b)_f)^2} \leq \sqrt{2c'} \frac{\lambda^d(V_{\vec{F}},V_{f})}{1-\lambda},
\]

with \(c' = \sup_{b \in \text{Im}(A)} \frac{W(b)\max_{e \in V_{\vec{F}}} \sqrt{2\lambda V(v)\sqrt{V_{f}}}}{\min\{d(b)_u,d(b)_v\}\min_{e \in V_{\vec{F}}} \sqrt{d(b)_e}}\).

5. Bias-variance and scale-free algorithms

Let us consider the network flow problem defined in the previous section, for a certain external flow \(b \in \mathbb{R}^V\) such that \(1^Tb = 0\). Let \(Z \subseteq V\), and choose \(p \in \mathbb{R}^V\) such that \(1^Tp = 0\) and such that \(p\) is supported on \(Z\), namely, \(p_v \neq 0\) if and only if \(v \in Z\). Assume that we perturb the external flow \(b\) by adding \(p\). We want to address the following question: given knowledge of the solution \(x^*(b)\) for the unperturbed problem, what is a computationally efficient algorithm to compute the solution \(x^*(b+p)\) of the perturbed problem? The main idea that we aim to exploit is that the decay of correlation property established in Theorem 8 implies that a localized perturbation of the external flow affects more the components of \(x^*(b)\) that are close to the perturbed sites. Hence, only a subset of the components of the
solution around $Z$ needs to be updated to meet a prescribed level of error precision, yielding savings on the computational cost.

To formalize this idea, let $\mathcal{G}' = (V', \mathcal{E}')$ be a subgraph of $\mathcal{G} = (V, \mathcal{E})$ such that $Z \subseteq V'$. Let $\tilde{G}' = (V', \tilde{E}')$ be the undirected graph associated to $\mathcal{G}'$ (see Remark 1), and assume that $G'$ is connected. Define $V'^C := V \setminus V'$ and $\tilde{E}'^C := \tilde{E} \setminus \tilde{E}'$. We now introduce a local algorithm to approximately compute $x^*(b + p)$ that only updates the components of $x^*(b)$ on $\tilde{E}'$.

5.1 Localized algorithms

As a general-purpose algorithm for constrained convex optimization, we consider the canonical projected gradient descent algorithm. The same argument about localization that we are about to present can analogously be developed for other optimization procedures. Recall that a single iteration of the projected gradient descent algorithm to compute $x^*(b)$ is the map defined as $T_b(x) := \arg\min\{\|u - (x - \eta \nabla f(u))\| : u \in \mathbb{R}^E, Au = b\}$, where $\eta > 0$ is a given step size. Let each function $f_e$ be $\alpha$-strongly convex and $\beta$-smooth, i.e., $\alpha \leq \frac{\partial^2 f_e(x)}{\partial x^2} \leq \beta$, for each $x \in \mathbb{R}$. A classical result yields that the projected gradient descent with step size $\eta = \frac{1}{\beta}$ converges to the optimal solution of the problem, namely, $\lim_{t \to \infty} T_b^t(x) = x^*(b)$ for any starting point $x \in \mathbb{R}^E$, where $T_b$ defines the $t$-th iteration of the algorithm. In the $\ell_2$-norm, the convergence rate is given by $\|T_b^t(x) - x^*(b)\| \leq e^{-t/(2Q)}\|x - x^*(b)\|$, where $Q = \beta/\alpha$ is the so-called condition number (see Bubeck (2015)[Theorem 3.6], for instance). We define the localized projected gradient descent on $\mathcal{G}'$ as follows (recall the notation from Remark 1).

**Definition 11 (Localized projected gradient descent)** Given $x \in \mathbb{R}^E$ such that we have $A_{V'C, E'C}x_{E'C} = b_{V'C}$, the localized projected gradient descent on $\mathcal{G}'$ with step size $\eta > 0$ is $T_b^t(x) := \arg\min\{\|u - (x - \eta \nabla f(u))\| : Au = b, u_{E'C} = x_{E'C}\}$.

Only the components of $x$ supported on $\tilde{E}'$ are updated by $T_b^t$, while the components on $\tilde{E}'^C$ stay fixed, playing the role of boundary conditions: for $e \in \tilde{E}'^C$ we have $T_b^t(x)_e = x_e$. For this reason, the map $T_b^t$ is defined only for the points $x \in \mathbb{R}^E$ whose coordinates outside $\tilde{E}'$ are consistent with the constraint equations. The algorithm that we propose to compute $x^*(b + p)$ given knowledge of $x^*(b)$ is easily described: it amounts to running for $t$ times the localized projected gradient descent on $\tilde{G}'$ with "frozen" boundary conditions $x^*(b)_{E'C}$ (and step size $\eta = 1/\beta$), namely, $T_{b^t+p}^t(x^*(b))$. Clearly, $x^*(b)$ satisfies the flow conservation constraints on $\tilde{E}'^C$, by definition.

5.2 Error analysis: bias-variance decomposition

We now provide estimates for the error committed by the localized projected gradient descent as a function of the subgraph $\tilde{G}'$ and the running time $t$. The key ingredient is the decay of correlation property for the network flow problem established in Theorem 8.

---

2. Although we present an application for expander graphs, the same idea behind our analysis works for other graph topologies upon using Theorem 7 to establish results in the same spirit of Theorem 8, possibly with other type of decays (i.e., not exponential), as discussed in Section 4.1.
Let us define the error committed by the localized projected gradient descent algorithm after $t \geq 1$ iterations as the vector in $\mathbb{R}^V$ given by $\text{Error}(p, \vec{G}', t) := x^*(b + p) - T^t_{b+p}(x^*(b))$. The analysis that we give is based on the following decomposition, that resembles the bias-variance decomposition in statistical analysis: $\text{Error}(p, \vec{G}', t) = \text{Bias}(p, \vec{G}') + \text{Variance}(p, \vec{G}', t)$, where $\text{Bias}(p, \vec{G}') := x^*(b + p) - \lim_{t \to \infty} T^t_{b+p}(x^*(b))$ and $\text{Variance}(p, \vec{G}', t) := \{\lim_{t \to \infty} T^t_{b+p}(x^*(b))\} - T^t_{b+p}(x^*(b))$. The bias term is algorithm-independent — any algorithm that converges to the optimal solution on the restricted problem yields the same bias — and it characterizes the error that we commit by localizing the optimization procedure per se, as a function of the subgraph $\vec{G}'$. On the other hand, the variance term depends on time and on the specific choice of the algorithm on $\vec{G}'$.

Let define the inner boundary of $\vec{G}'$ as $\Delta(\vec{G}') := \{v \in V' : \mathcal{N}(v) \cap V'^C \neq \emptyset\}$, which represents the subset of vertices in $V'$ that have at least one vertex neighbor outside $V'$ (in the undirected graph $G$). Let $B \in \mathbb{R}^{V \times V}$ be the vertex-vertex adjacency matrix of the undirected graph $G = (V, E)$, which is the symmetric matrix defined as $B_{uv} := 1$ if $\{u, v\} \in E$, $B_{uv} := 0$ otherwise. Being real and symmetric, the matrix $B$ has $n := |V|$ real eigenvalues which we denote by $\mu_n \leq \mu_{n-1} \leq \cdots \leq \mu_2 \leq \mu_1$. Let $\mu := \max\{|\mu_2|, |\mu_n|\}$ be the second largest eigenvalue in magnitude of $B$. The next theorem yields bounds for the bias and variance error terms in the $\ell_2$-norm. The bound for the bias decays exponentially with respect to the graph-theoretical distance (i.e., the distance in the unweighted graph $G$) between the inner boundary of $\vec{G}'$, i.e., $\Delta(\vec{G}')$, and the region where the perturbation $p$ is supported, i.e., $Z \subseteq V$. The rate is governed by the eigenvalue $\mu$, the condition number $Q$, and the maximum/minimum degree of the graph. The bound for the variance decays exponentially with respect to the running time, with rate proportional to $1/Q$. The proof of this theorem is given in Appendix E, and the key ingredient is the decay of correlation property established in Theorem 8.

**Theorem 12 (Error localized algorithm)** Let $k_-$ and $k_+$ be, respectively, the minimum and maximum degree of $G$. Let $\rho := \frac{Qk_+}{k_-} - 1 + \frac{Q}{k_-} \mu$. If $\rho < 1$, then

$$\| \text{Bias}(p, \vec{G}') \| \leq \| p \| \gamma \rho^{|\Delta(\vec{G}')|Z} \frac{1}{(1 - \rho)^2} 1_{\vec{G}' \neq \vec{G}}$$

$$\| \text{Variance}(p, \vec{G}', t) \| \leq \| p \| c e^{-t/(2Q)} \frac{1}{1 - \rho},$$

with $\gamma := c(1 + c\sqrt{k_- - 1}), c := \sqrt{\frac{2k_+}{k_-}}Q$. We have

$$\| \text{Error}(p, \vec{G}', t) \| \leq \| \text{Bias}(p, \vec{G}') \| + \| \text{Variance}(p, \vec{G}', t) \|.$$

Note that the constants appearing in the bounds in Theorem 12 do not depend on the choice of the subgraph $\vec{G}'$ of $\vec{G}$, but depend only on $\mu$, $Q$, $k_+$, and $k_-$ (as the proofs in Appendix E attest, a more refined analysis can yield better constants that do depend on the choice of $\vec{G}'$). In particular, the same constants apply for the analysis of the global algorithm, i.e., the projected gradient descent applied to the entire graph $\vec{G}$. In this case, the bias term clearly equals 0, so that the error equals the variance (hence the indicator function $1_{\vec{G}' \neq \vec{G}}$ in Theorem 12).
In the next section we show that the bias introduced by the localization procedure can be exploited to lower the computational complexity that is associated to the variance term. This is the key idea that allows us to prove dimension-free computational complexity for the localized algorithm.

5.3 Bias-variance trade-off. Dimension-free complexity

The error estimates established in Theorem 12 allow us to prove that the localized projected gradient descent is scale-free, in the sense that it is guaranteed to meet a prescribed level of error accuracy $\varepsilon > 0$ with a computational complexity that does not depend on the dimension of $\vec{G}$. To illustrate this fact, let $G = (V, E)$ be a $k$-regular graph such that the second largest eigenvalue in magnitude of its vertex-vertex adjacency matrix is bounded away from $k$ as a function of the dimension of $G$: namely, $\mu \leq \gamma < k$, where $\gamma$ is a universal constant that does not depend on the size $|V|$, nor on the size $|E|$. This is the same as saying that $G$ comes from a family of $k$-regular expander graphs Hoory et al. (2006). Define $\vec{G} = (V, \vec{E})$ by assigning an arbitrary orientation to the edges of $G$. Assume that the following holds: $\rho = Q - 1 + \frac{Q}{k} \mu < 1$, where recall that $Q = \beta / \alpha$ is the condition number.

For the sake of simplicity, we introduce a collection of subgraphs of $\vec{G}$ that are centered on a given vertex and are parametrized by their radii, where we stress that the notion of distance is with respect to the undirected graph $G$. Namely, fix a vertex $v \in V$, let $V_r := \{ w \in V : d(v, w) \leq r \}$ denote the ball of radius $r > 0$ around vertex $v \in V$, and let $\vec{G}_r := (V_r, \vec{E}_r)$ be the subgraph of $\vec{G}$ that has vertex set $V_r$ and induced edge set $\vec{E}_r$. Let $r_{\max} := \max \{ d(v, w) : w \in V \}$. Consider a perturbation vector $p \in \mathbb{R}^V$ that is supported on $Z := V_z$, for a fixed $z \geq 1$. See Figure 2.

![Figure 2: Representation of a directed graph $\vec{G} = (V, \vec{E})$, subgraph $\vec{G}_z = (V_z, \vec{E}_z)$ with $z = 1$ (vertices and edges fully contained in the inner red circle), and subgraph $\vec{G}_r = (V_r, \vec{E}_r)$ with $r = 3$ (vertices and edges fully contained in the outer blue circle). In this case, $r_{\max} = 4$.](image-url)

If we run the localized algorithm on $\vec{G}_r$, with $r > z$, for $t$ time steps, then Theorem 12 yields the following estimate (here $d(\Delta(\vec{G}_r), Z) = r - z$ and $1_{\vec{G}_r \neq \vec{G}} = 1_{r < r_{\max}}$): $\| \text{Error}(p, \vec{G}_r, t) \| \leq \| p \| \nu_{\text{bias}} e^{-\xi_{\text{bias}} r} 1_{r < r_{\max}} + \| p \| \nu_{\text{var}} e^{-\xi_{\text{var}} t}$, with $\nu_{\text{bias}} := \gamma (1 - \rho)^2 \rho^2$, $\xi_{\text{bias}} := \log \frac{1}{\rho} > 0$, $\nu_{\text{var}} := \frac{c}{1 - \rho}$, $\xi_{\text{var}} := \frac{1}{2Q} > 0$, with $\gamma := c(1 + c\sqrt{k} - 1)$, $c := \sqrt{2Q}/\sqrt{k}$. Let $\kappa(\vec{G}_r, t)$
be the computational complexity required to run the localized projected gradient descent algorithm on $\tilde{G}_r$ for $t$ time steps. A rough estimate for the asymptotic behavior of $\kappa(\tilde{G}_r, t)$ is easily derived as follows (more refined estimates can be made, but we do not need them for the sake of the present argument). If $A_r := A_{V_r, E_r}$ denotes the vertex-edge adjacency matrix associated to $\tilde{G}_r$, and $f_r := \sum_{e \in E_r} f_e$ is the restriction of the function $f$ to the edges in $\tilde{G}_r$, it is easy to check that a single iteration of the localized projected gradient descent algorithm on $\tilde{G}_r$ reads $T_{b+\rho}^{(r)}(x)_{E_r} = (I - A_r^T (A_r A_r^T)^+ A_r) (x_{E_r} - \eta \nabla f_r(x_{E_r})) + A_r^T (A_r A_r^T)^+ (b_v + p_{V_r} - A_{V_r, E_C} x_{E_C}), T_{b+\rho}^{(r)}(x)_{E_C} = x_{E_C}$, for any $x \in \mathbb{R}^E$ such that $A_{V_C, E_C} x_{E_C} = b_{V_C}$. The exact computation of the matrix $(A_r A_r^T)^+$ has an asymptotic complexity that scales like $O(|V_r|^\omega)$ as a function of $r$, where $\omega > 2$ is the so-called matrix multiplication constant. As each matrix-vector multiplication has a cost of $O(|V_r|^2)$, then $\kappa(\tilde{G}_r, t)$ scales like $O(|V_r|^\omega + |V_r|^2 t)$. For the sake of simplicity, consider $O(|V_r|^\omega t)$. To estimate the complexity of the local algorithm, we need to bound the growth of $|V_r|$ as a function of $r$. In a $k$-regular graph, we clearly have $|V_r| \leq k^r$ (which is realistic for expander graphs, as they are locally tree-like) so that $\kappa(\tilde{G}_r, t)$ grows at most as $O(e^{(\omega \log k) r t})$.

We are now in the position to appreciate the computational savings that the localized algorithm offers over the global algorithm (i.e., the projected gradient descent on $\tilde{G}$). Assume that $\tilde{G}$ is an infinite network with $r_{\text{max}} = \infty$. In this case, the computational complexity of the global algorithm is clearly infinity, as the global algorithm updates the components of the solution at every edge of the entire network. On the other hand, the complexity of the localized projected gradient descent algorithm is finite. This can be seen if we seek, for example, for the minimal radius $r$ and time $t$ such that $\nu_{\text{bias}} e^{-\xi \kappa(t)} \leq \frac{\epsilon}{2}$ and $\nu_{\text{var}} e^{-\xi \kappa(t)} \leq \frac{\epsilon}{2}$. Clearly, these constraints guarantee that $\|\text{Error}(p, G_r, t)\| \leq \epsilon$, and it is easy to see that both the minimal $t$ and the minimal $r$ that satisfy the above inequalities scale like $O(\log(\|p\|/\epsilon))$, so that the complexity of the localized algorithm scales like $O((\|p\|/\epsilon)^{\omega \log k} \log(\|p\|/\epsilon))$, where the constants involved do not dependent of the dimension of the graph $\tilde{G}$, but depend only on $\mu$, $Q_1$, and $k$.

The decay of correlation property exhibited by the network flow problem allowed us to show that the bias introduced by localizing the optimization problem to a subgraph $\tilde{G}_r$ saves us from the computational cost associated to the variance term, which corresponds to running the gradient descent algorithm on $\tilde{G}_r$ for $t$ time steps. In fact, a finer analysis shows that one can exploit the bias-variance trade-off to optimally tune the algorithm, i.e., to find a radius $r(\epsilon)$ and a time $t(\epsilon)$ that minimize the computational complexity $\kappa(\tilde{G}_{r(\epsilon)}, t(\epsilon))$ which is required to reach the prescribed level of error accuracy $\epsilon$. These ideas suggest a general framework to study the trade-off between accuracy and complexity for local algorithms in network optimization.

---

3. The same rationale behind the argument that we make applies if we consider approximate algorithms that are taylor-made to take advantage of the Laplacian structure of the matrix $A_r A_r^T$ and yield much better computational complexity to $\delta$-compute $(A_r A_r^T)^+$, of the order of $\tilde{O}(|E_r| \log |V_r| \log(1/\delta))$, see Koutis et al. (2011).
6. Conclusions

The main contribution of this paper is to derive a general analogy between natural concepts in probability and statistics (i.e., notions of correlation among random variables, decay of correlation, and bias-variance decomposition and trade-off) and similar notions that can be introduced in optimization. In this paper we have proposed general notions of correlation that are based on the sensitivity of optimal points. We have illustrated how decay of correlation (locality) can be established in a canonical network optimization problem (min-cost network flow), and how it can be used to design a localized version of a canonical algorithm (projected gradient descent) in an instance of reoptimization (warm-start after localized perturbations). In principle, the framework that we propose can be applied to any optimization problem and to any algorithm. The key point is to establish decay of correlation for the problem at hand, deriving results that are analogous to the ones established in Section 4.2. In the case of the min-cost network flow problem, we have shown that establishing locality reduces to bounding the discrete derivative of the Green’s function of the diffusion random walk, as described in Theorem 7. We proved exponential decay for the correlation in expander graphs, and we conjectured polynomial decay in grid-like topologies. Once results on locality are established for the particular problem at hand, they translate into a bound for the bias term of the error decomposition of localized algorithms that we give in Theorem 12. The analysis of the variance term, on the other hand, depends on the algorithm that one wants to localize. This part is not technically difficult, as it amounts to analyzing the performance of the chosen algorithm when applied to a subgraph with frozen boundary conditions, as it can be seen from the proof of Theorem 12 in Appendix E. Establishing locality in more general settings than the one here considered and extending the ideas here presented to cold-start scenarios (where one wants to compute the optimal solution of an optimization problem starting from possibly any initial condition) remain open questions for future investigation.

References


**Appendix A. Hadamard’s global inverse theorem**

In this section we provide the proof of Theorem 2 in Section 2. The proof relies on Hadamard’s global inverse function theorem, which we first state. Recall that a function from \( \mathbb{R}^m \) to \( \mathbb{R}^m \) is said to be \( C^k \) if it has continuous derivatives up to order \( k \). A function is said to be a \( C^k \) diffeomorphism if it is \( C^k \), bijective, and its inverse is also \( C^k \). The following important result characterizes when a \( C^k \) function is a \( C^k \) diffeomorphism.

**Theorem 13 (Hadamard’s global inverse function theorem)** Let \( \Psi \) be a \( C^k \) function from \( \mathbb{R}^m \) to \( \mathbb{R}^m \). Then, \( f \) is a \( C^k \) diffeomorphism if and only if the following two conditions hold:

1. The determinant of the differential of \( \Psi \) is different from zero at any point, namely, \( \frac{d}{dz}\Psi(z) \neq 0 \) for any \( z \in \mathbb{R}^m \).

2. The function \( \Psi \) is norm coercive, namely, for any sequence of points \( z_1, z_2, \ldots \in \mathbb{R}^m \) with \( \|z_k\| \to \infty \) it holds \( \|\Psi(z_k)\| \to \infty \) (for any choice of the norm \( \|\cdot\| \), as norms are equivalent in finite dimension).

**Proof** See Wu and Desoer (1972)[Corollary of Lemma 2], or also Krantz and Parks (2002) for a more general form of this inverse function to study diffeomorphisms on manifolds. □

The following corollary is the backbone behind Theorem 2.

**Lemma 14 (Diffeomorphism for Lagrangian multipliers map)** Let \( f : \mathbb{R}^n \to \mathbb{R} \) be a strongly convex function, twice continuously differentiable. Let \( A \in \mathbb{R}^{m \times n} \) be a given matrix. Define the function \( \Phi \) from \( \mathbb{R}^n \times \mathbb{R}^m \) to \( \mathbb{R}^n \times \mathbb{R}^m \) as

\[
\Phi(x, \nu) := \begin{pmatrix} \nabla f(x) + A^T \nu \\ Ax \end{pmatrix},
\]

for any \( x \in \mathbb{R}^n \), \( \nu \in \mathbb{R}^m \). Then, the restriction of the function \( \Phi \) to \( \mathbb{R}^n \times \text{Im}(A) \) is a \( C^1 \) diffeomorphism.


**Proof** Let us interpret $\Phi$ as the representation of a transformation $\mathcal{T}$ in the standard basis of $\mathbb{R}^n \times \mathbb{R}^m$. Recall the orthogonal decomposition $\mathbb{R}^m = \text{Im}(A) \oplus \text{Ker}(A^T)$. Let the vectors $u_1, \ldots, u_r \in \mathbb{R}^m$ form an orthogonal basis for $\text{Im}(A)$, where $r$ is the rank of $A$, and let the vectors $v_1, \ldots, v_{m-r} \in \mathbb{R}^m$ form an orthogonal basis for $\text{Ker}(A^T)$. Define the orthogonal matrix $Z = [u_1, \ldots, u_r, z_1, \ldots, z_{m-r}]$, which represents a change of basis in $\mathbb{R}^m$. As we have

$$\Phi(x, \nu) = \left( \nabla f(x) + A^T ZZ^T \nu \right),$$

then the transformation $\mathcal{T}$ is represented in the standard basis for $\mathbb{R}^n$ and in the basis $Z$ for $\mathbb{R}^m$ by the following map

$$\tilde{\Phi}(x, \tilde{\nu}) := \left( \nabla f(x) + \tilde{A}^T \tilde{\nu} \right),$$

where $\tilde{A} := Z^T A$. In fact,

$$\tilde{\Phi}(x, Z^T \nu) = \left( \begin{array}{cc} I & 0 \\ 0 & Z^T \end{array} \right) \Phi(x, \nu),$$

where $I \in \mathbb{R}^{n \times n}$ is the identity matrix, and $0 \in \mathbb{R}^{n \times m}$ is the all-zero matrix. As $A^T Z = [A^T u_1, \ldots, A^T u_r, 0_{n \times (m-r)}]$,

$$\tilde{A} = (A^T Z)^T = \left[ \begin{array}{c} B \\ 0_{(m-r) \times n} \end{array} \right],$$

where $B := [u_1, \ldots, u_r]^T A \in \mathbb{R}^{r \times n}$. Therefore, the restriction of the transformation $\mathcal{T}$ to the invariant subspace $\mathbb{R}^n \times \text{Im}(A)$ is represented in the standard basis for $\mathbb{R}^n$ and in the basis $\{u_1, \ldots, u_r\}$ for $\text{Im}(A)$ by the following map

$$\Psi(x, \xi) := \left( \nabla f(x) + B^T \xi \right).$$

As the function $f$ is twice continuously differentiable, clearly the function $\Psi$ is continuously differentiable, i.e., $C^1$. We check that the two conditions of Theorem 13 are met.

The differential of $\Psi$ evaluated at $(x, \xi) \in \mathbb{R}^n \times \mathbb{R}^r$ is

$$J(x, \xi) := \left( \begin{array}{cc} \nabla^2 f(x) & B^T \\ B & 0 \end{array} \right).$$

As $f$ is strongly convex, $\nabla^2 f(x)$ is positive definite so invertible. Then, the determinant of the Jacobian can be expressed as $|J(x, \xi)| = |\nabla^2 f(x)|| - B \nabla^2 f(x)^{-1} B^T|$. As $B$ has full row rank by definition, $B \nabla^2 f(x)^{-1} B^T$ is positive definite and we clearly have $|J(x, \xi)| \neq 0$.

To prove that the function $\Psi$ is norm coercive, let us choose $\| \cdot \|$ to be the Euclidean norm and consider a sequence $(x_1, \xi_1), (x_2, \xi_2), \ldots \in \mathbb{R}^n \times \mathbb{R}^r$ with $\|(x_k, \xi_k)\| \rightarrow \infty$. As for any $x \in \mathbb{R}^n, \xi \in \mathbb{R}^r$ we have $\| (x, \xi) \|^2 = \|x\|^2 + \|\xi\|^2$, clearly for the sequence to go to infinity one of the following two cases must happen:
(a) \( \|x_k\| \to \infty \);
(b) \( \|x_k\| \leq c \) for some \( c < \infty \), \( \|\xi_k\| \to \infty \).

Before we consider these two cases separately, let us note that, for any \( x \in \mathbb{R}^n, \xi \in \mathbb{R}^r \),

\[
\|\Psi(x, \xi)\|^2 = \|\nabla f(x) + B^T \xi\|^2 + \|Bx\|^2.
\]  

(2)

Let \( \alpha > 0 \) be the strong convexity parameter, and recall the following definition of strong convexity, for any \( x, y \in \mathbb{R}^n \),

\[
(\nabla f(x) - \nabla f(y))^T (x - y) \geq \alpha \|x - y\|^2.
\]  

(3)

(a) Assume \( \|x_k\| \to \infty \). Let \( P_\parallel \) be the projection operator on \( \text{Im}(B^T) \), \( P := B^T (BB^T)^{-1}B \), and let \( P_\perp = I - P_\parallel \) be the projection operator on \( \text{Ker}(B) \), the orthogonal complement of \( \text{Im}(B^T) \). As for any \( x \in \mathbb{R}^n \) we have the decomposition \( x = P_\parallel x + P_\perp x \) with \( (P_\parallel x)^T P_\perp x = 0 \), clearly \( \|x\|^2 = \|P_\parallel x\|^2 + \|P_\perp x\|^2 \). So, the condition \( \|x_k\| \to \infty \) holds only if one of the two cases happens:

(i) \( \|P_\parallel x_k\| \to \infty \);

(ii) \( \|P_\parallel x_k\| \leq c \) for some \( c < \infty \), \( \|P_\perp x_k\| \to \infty \).

Consider (i) first. Let \( x \in \mathbb{R}^n \) so that \( P_\parallel x \neq 0 \). As \( BP_\parallel = 0 \), from (2) we have, for \( \xi \in \mathbb{R}^r \), \( \|\Psi(x, \xi)\|^2 \geq \|Bx\|^2 = \|BP_\parallel x\|^2 \), for which \( \min_{y \in \mathbb{R}^n, y \neq 0} \frac{y^T B^T B y}{\|y\|^2} \|P_\parallel x\|^2 = \lambda \|P_\parallel x\|^2 \) is a lower bound, where \( \lambda \) is the minimum eigenvalue of \( B^T B \) among those corresponding to the eigenvectors spanning the subspace \( \text{Im}(B^T) \). Clearly, if \( \lambda \neq 0 \) (notice \( \lambda \geq 0 \) by definition) then the above yields that \( \|\Psi(x_k, \xi_k)\| \to \infty \) whenever \( \|P_\parallel x_k\| \to \infty \). To prove this, assume by contradiction that \( \lambda = 0 \). Then, there exists \( y \in \mathbb{R}^n \) satisfying \( y \in \text{Im}(B^T) \), \( y \neq 0 \), such that \( B^T B y = \lambda y = 0 \). As \( B^T \) has full column rank, the latter is equivalent to \( By = 0 \) so that \( P_\perp y = y \neq 0 \), which contradicts the hypothesis that \( y \in \text{Im}(B^T) \).

Consider now the case (ii). Decomposing the gradient on \( \text{Im}(B^T) \) and its orthogonal subspace, from (2) we have, for any \( x \in \mathbb{R}^n, \xi \in \mathbb{R}^r \), \( \|\Psi(x, \xi)\|^2 \geq \|P_\parallel \nabla f(x) + P_\parallel \nabla f(x) + B^T \xi\|^2 = \|P_\parallel \nabla f(x)\|^2 + \|P_\perp \nabla f(x) + B^T \xi\|^2 \), so that \( \|\Psi(x, \xi)\| \geq \|P_\perp \nabla f(x)\| \). Choosing \( y = P_\parallel x \) in (3) we have \( (P_\perp \nabla f(x) - P_\parallel \nabla f(P_\parallel x))^T P_\perp x \geq \alpha \|P_\perp x\|^2 \), and applying Cauchy-Schwarz we get, for any \( x \) with \( P_\perp x \neq 0 \), \( \|P_\perp \nabla f(x)\| \geq \alpha \|P_\perp x\| - \|P_\perp \nabla f(P_\parallel x)\| \). By assumption \( f \) is twice continuously differentiable, so \( \nabla f \) is continuous and it stays bounded on a bounded domain. Hence, we can conclude that \( \|\Psi(x_k, \xi_k)\| \to \infty \) if \( \|P_\perp x_k\| \to \infty \) with \( (P_\parallel x_k)_{k \geq 1} \) bounded.

(b) Assume \( \|\xi_k\| \to \infty \) and \( (x_k)_{k \geq 1} \) bounded. For any \( \xi \in \mathbb{R}^r, \xi \neq 0 \), we have

\[
\|B^T \xi\|^2 = \frac{\xi^T B B^T \xi}{\|\xi\|^2} \geq \min_{y \in \mathbb{R}^n, y \neq 0} \frac{y^T B B^T y}{\|y\|^2} \|\xi\|^2 = \lambda_{\text{min}} \|\xi\|^2,
\]

where \( \lambda_{\text{min}} \) is the minimum eigenvalue of \( B B^T \), which is strictly positive as \( B B^T \) is positive definite by the assumption that \( B \) has full row rank. From (2) we have

\[
\|\Psi(x, \xi)\| \geq \|\nabla f(x) + B^T \xi\| \geq \|B^T \xi\| - \|\nabla f(x)\| \geq \sqrt{\lambda_{\text{min}} \|\xi\| - \|\nabla f(x)\|},
\]

that, by continuity of \( \nabla f \), shows that \( \|\Psi(x_k, \xi_k)\| \to \infty \) if \( \|\xi_k\| \to \infty \) and \( (x_k)_{k \geq 1} \) is bounded.
We are finally ready to state the proof of Theorem 2.

**Proof** [Proof of Theorem 2] The Lagrangian of the optimization problem is the function $L$ from $\mathbb{R}^V \times \mathbb{R}^F$ to $\mathbb{R}$ defined as $L(x, \nu) := f(x) + \sum_{a \in F} \nu_a (A^T_a x - b_a)$, where $A^T_a$ is the a-th row of the matrix $A$ and $\nu = (\nu_a)_{a \in F}$ is the vector formed by the Lagrangian multipliers.

Let us define the function $\Phi$ from $\mathbb{R}^L \times \mathbb{R}^L$ definite. As $\text{Im}(L)$, we show that $\text{Ker}(L)$ exists for the unique minimizer $x^*(b(\varepsilon))$ so that

$$\Phi(x, \nu) := \left( \begin{array}{c} \nabla_x L(x, \nu) \\ A x \end{array} \right) = \left( \begin{array}{c} \nabla f(x) + A^T \nu \\ A x \end{array} \right).$$

For any fixed $\varepsilon \in \mathbb{R}$, as the constraints are linear, the Lagrange multiplier theorem says that for the unique minimizer $x^*(b(\varepsilon))$ there exists $\nu'(b(\varepsilon)) \in \mathbb{R}^F$ so that

$$\Phi(x^*(b(\varepsilon)), \nu'(b(\varepsilon))) = \left( \begin{array}{c} 0 \\ b(\varepsilon) \end{array} \right).$$

As $A^T(\nu + \mu) = A^T \nu$ for each $\mu \in \text{Ker}(A^T)$, the set of Lagrangian multipliers $\nu'(b(\varepsilon)) \in \mathbb{R}^F$ that satisfies (4) is a translation of the null space of $A^T$. We denote the unique translation vector by $\nu^*(b(\varepsilon)) \in \text{Im}(A)$. By Hadamard’s global inverse function theorem, as shown in Lemma 14, the restriction of the function $\Phi$ to $\mathbb{R}^V \times \text{Im}(A)$ is a $C^1$ diffeomorphism, namely, it is continuously differentiable, bijective, and its inverse is also continuously differentiable. In particular, this means that the functions $x^* : b \in \text{Im}(A) \rightarrow x^*(b) \in \mathbb{R}^V$ and $\nu^* : b \in \text{Im}(A) \rightarrow \nu^*(b) \in \text{Im}(A)$ are continuously differentiable along the subspace $\text{Im}(A)$.

Differentiating both sides of (4) with respect to $\varepsilon$, we get

$$\left( \begin{array}{cc} H & A^T \\ A & 0 \end{array} \right) \left( \begin{array}{c} x' \\ \tilde{\nu} \end{array} \right) = \left( \begin{array}{c} 0 \\ \frac{d b(\varepsilon)}{d \varepsilon} \end{array} \right),$$

where $H := \nabla^2 f(x^*(b(\varepsilon)))$, $x' := \frac{dx^*(b(\varepsilon))}{d \varepsilon}$, $\tilde{\nu} := \frac{d \nu^*(b(\varepsilon))}{d \varepsilon}$. As the function $f$ is strongly convex, the Hessian $\nabla^2 f(x)$ is positive definite for every $x \in \mathbb{R}^V$, hence it is invertible for every $x \in \mathbb{R}^V$. Solving the linear system for $x'$ first, from the first equation $H x' + A^T \tilde{\nu} = 0$ we get $x' = -H^{-1} A^T \tilde{\nu}$. Substituting this expression in the second equation $A x' = b(\varepsilon)$, we get $L \tilde{\nu} = -\frac{d b(\varepsilon)}{d \varepsilon}$, where $L := A H^{-1} A^T$. The set of solutions to $L \tilde{\nu} := -\frac{d b(\varepsilon)}{d \varepsilon}$ can be expressed in terms of the pseudoinverse of $L$ as follows $\{ \tilde{\nu} \in \mathbb{R}^F : L \tilde{\nu} = -\frac{d b(\varepsilon)}{d \varepsilon} \} = \text{Ker}(L)$. We show that $\text{Ker}(L) = \text{Ker}(A^T)$. We show that $L \nu = 0$ implies $A^T \nu = 0$, as the opposite direction trivially holds. In fact, let $A' := A \sqrt{H^{-1}}$, where $\sqrt{H^{-1}}$ if the positive definite matrix that satisfies $\sqrt{H^{-1}} \sqrt{H^{-1}} = H^{-1}$. The condition $L \nu = A' A^T \nu = 0$ is equivalent to $A' A^T \nu \in \text{Ker}(A')$. At the same time, clearly, $A' A^T \nu \in \text{Im}(A^T)$. However, $\text{Ker}(A')$ is orthogonal to $\text{Im}(A^T)$, so it must be $A' A^T \nu = 0$ which implies $A^T \nu = 0$ as $\sqrt{H^{-1}}$ is positive definite. As $\text{Im}(L^+) = \text{Ker}(L)^\perp = \text{Ker}(A^T)^\perp = \text{Im}(A)$, we have that $\tilde{\nu} = -L^+ \frac{d b(\varepsilon)}{d \varepsilon}$ is the unique solution to $L \tilde{\nu} = -\frac{d b(\varepsilon)}{d \varepsilon}$ that belongs to $\text{Im}(A)$. Substituting this expression into $x' = -H^{-1} A^T \tilde{\nu}$, we finally get $x' = H^{-1} A^T L^+ \frac{d b(\varepsilon)}{d \varepsilon}$. The proof follows as $\Sigma(b) = H^{-1}$. ■
Appendix B. Correlation

In this section we provide the proofs of Proposition 4 and Lemma 5 in Section 3.

Proof [Proof of Proposition 4] Let \( \mathcal{I} \) be a finite set, and let \( Z \in \mathbb{R}^\mathcal{I} \) be a Gaussian random vector with mean \( \rho \in \mathbb{R}^\mathcal{I} \) and covariance \( \Upsilon \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}} \), not necessarily invertible. Let \( \mathcal{V} \subseteq \mathcal{I} \) be not empty, and \( \mathcal{F} := \mathcal{I} \setminus \mathcal{V} \) not empty. Recall that if \( z_\mathcal{F} \in \text{Im}(\Upsilon_{\mathcal{F},\mathcal{F}}) \), where \( \Upsilon_{\mathcal{F},\mathcal{F}} \) is the submatrix of \( \Upsilon \) indexed by the rows and columns referring to \( \mathcal{F} \), then the conditional mean of \( Z_\mathcal{V} \) given \( Z_\mathcal{F} = z_\mathcal{F} \) is given by (see Albert (1972)[Theorem 9.2.1], for instance): 
\[
\mathbb{E}[Z_\mathcal{V}|Z_\mathcal{F} = z_\mathcal{F}] = \rho_\mathcal{V} + \Upsilon_{\mathcal{V},\mathcal{F}}(\Upsilon_{\mathcal{F},\mathcal{F}})^{-1}(z_\mathcal{F} - \rho_\mathcal{F}).
\] 
The statement of the proposition follows immediately if we consider the Gaussian vector \( Z := (X,A) \) with \( Z_\mathcal{V} := X \), \( Z_\mathcal{F} := AX \), upon noticing that \( \rho_\mathcal{V} = \mathbb{E}[X] = \mu \), \( \rho_\mathcal{F} = \mathbb{E}[AX] = A\mu \), \( \Upsilon_{\mathcal{V},\mathcal{F}} = \text{Cov}(X,A) = \mathbb{E}[X(AX)^T] = \mathbb{E}[XX^TA^T] = \Sigma A^T \), and \( \Upsilon_{\mathcal{F},\mathcal{F}} = \text{Cov}(AX,A) = \mathbb{E}[AXXX^T] = \Sigma A^T \).

If \( \Sigma \) is invertible, i.e., positive definite, and \( A \) is full rank, then also \( \Sigma A^T \) is positive definite and invertible, so \( (\Sigma A^T)^{-1} = (\Sigma A^T)^{-1} \).

Proof [Proof of Lemma 5] Given \( b \in \mathbb{R}^B \), note that \( x^*_I(b) \) corresponds to the components labeled by \( I \) of the solution \( x^*(b) \) of the optimization problem (1), with \( A := (0,I) \), where \( 0 \) is the all-zero matrix in \( \mathbb{R}^{B \times I} \) and \( I \) is the identity matrix in \( \mathbb{R}^{B \times B} \), and \( x = (x_I,x_B)^T \). As the matrix \( A \) is clearly full row rank, and \( H = \nabla^2 f(x^*_I(b)) = \nabla^2 f(x^*(b)) \), Corollary 3 yields
\[
\frac{dx^*_I(b)}{db} = \Sigma A^T (\Sigma A^T)^{-1} = \left( \Sigma_I B (\Sigma_B,B)^{-1} \right),
\]
so that \( \frac{dx^*_I(b)}{db} = \Sigma_I B (\Sigma_B,B)^{-1} \). The matrix identity \( \Sigma_I B (\Sigma_B,B)^{-1} = -(H_{I,I})^{-1}H_{I,B} \) is standard textbook material. However, to show the statement involving the matrix \( H \), we proceed from first principles, by applying the first order optimality condition to the restricted problem on \( I \). Note that \( x^*_I(x_B) \) is defined by the optimality conditions \( \frac{df(x_I^*(x_B)x_B)}{dx_I} = 0 \).

Differentiating with respect to \( x_B \), we get \( \frac{d^2 f(x_I^*(x_B)x_B)}{dx_I^2} \frac{dx_I^*(x_B)}{dx_B} + \frac{d^2 f(x_I^*(x_B)x_B)}{dx_B dx_I} = 0 \), or, equivalently, \( H_{I,I} \frac{dx_I^*(x_B)}{dx_I} = -H_{I,B} \). The proof is concluded by inverting the matrix \( H_{I,I} \), which is invertible as \( H \) is positive definite by assumption, so that any principal submatrix of it is also positive definite.

Appendix C. Graph Laplacians and random walks

This section is self-contained, and it provides several connections between graph Laplacians and random walks on weighted graphs. In particular, the connection between pseudoinverses of graph Laplacians and Green’s functions of random walks that we present in Lemma 18 below is the key result that allows us to derive spectral bounds to establish the decay of correlation property in Section 4, as showed in Appendix D.

Throughout, let \( G = (V,E,W) \) be a simple (i.e., no self-loops, and no multiple edges), connected, undirected, weighted graph, where to each edge \( \{v,w\} \in E \) is associated a non-negative weight \( W_{vw} = W_{wv} > 0 \), and \( W_{vw} = 0 \) if \( \{v,w\} \not\in E \). Let \( D \) be a diagonal matrix with entries \( d_v = D_{vv} = \sum_{w \in V} W_{vw} \) for each \( v \in V \). For each vertex \( v \in V \), let
$\mathcal{N}(v) := \{w \in V : \{v, w\} \in E\}$ be the set of node neighbors of $v$. In this section we establish several connections between the graph Laplacian $L := D - W$ and the random walk $X := (X_t)_{t \geq 0}$ with transition matrix $P := D^{-1}W$. Henceforth, for each $v \in V$, let $P_v$ be the law of a time homogeneous Markov chain $X_0, X_1, X_2, \ldots$ on $V$ with transition matrix $P$ and initial condition $X_0 = v$. Analogously, denote by $E_v$ the expectation with respect to this law. The hitting time to the site $v \in V$ is defined as $T_v := \inf\{t \geq 0 : X_t = v\}$. Let $\pi$ be the unique stationary distribution of the random walk, namely, $\pi^T P = \pi^T$. By substitution it is easy to check that $\pi_v := \frac{d_v}{\sum_{u \in V} d_u}$ for each $v \in V$. We adopt the notation $e_v \in \mathbb{R}^V$ to denote the vector whose only non-zero component equals 1 and corresponds to the entry associated to $v \in V$.

C.1 Restricted Laplacians and killed random walks

The connection between Laplacians and random walks that we present in Section C.2 below is established by investigating restricted Laplacians and killed random walks. Throughout this section, let $\bar{z} \in V$ be fixed. Let $\bar{V} := V \setminus Z$, $\bar{E} := E \setminus \{(u, v) \in E : u \in Z \text{ or } v \in Z\}$ and consider the graph $\bar{G} := (\bar{V}, \bar{E})$. Define $\bar{W}$ and $\bar{D}$ as the matrix obtained by removing the $\bar{z}$-th row and $\bar{z}$-th column from $W$ and $D$, respectively. Then, $\bar{L} := \bar{D} - \bar{W} \in \mathbb{R}^{\bar{V} \times \bar{V}}$ represents the so-called restricted Laplacian that is obtained by removing the $\bar{z}$-th row and $\bar{z}$-th column from $L$. Define $\bar{P} := \bar{D}^{-1}\bar{W} \in \mathbb{R}^{\bar{V} \times \bar{V}}$.

**Proposition 15** For $v, w \in \bar{V}, t \geq 0$, $\bar{P}_{vw}^t = P_v(X_t = w, T_{\bar{z}} > t)$.

The following proposition relates the inverse of the reduced Laplacian $\bar{L}$ with the Green function of the killed random walk, namely, the function $(u, v) \in \bar{V} \times \bar{V} \rightarrow \sum_{t=0}^{\infty} \bar{P}_{uv}^t$, and with the hitting times of the original random walk $X$.

**Proposition 16** For each $v, w \in \bar{V}$, we have

$$\bar{L}_{vw}^{-1} = \frac{1}{d_w} \sum_{t=0}^{\infty} \bar{P}_{vw}^t = \bar{L}_{vw}^{-1} P_v(T_w < T_{\bar{z}}),$$

and $\bar{L}_{ww}^{-1} = \frac{1}{d_w} E_v[\sum_{t=0}^{T_{\bar{z}}} 1_{X_t = w}]$.

**Proof** Let us first assume that $\bar{G}$ is connected. The matrix $\bar{P}$ is sub-stochastic as, clearly, if $v \notin \mathcal{N}(\bar{z})$ then $\sum_{w \in V} \bar{P}_{vw} = 1$, while if $v \in \mathcal{N}(\bar{z})$ then $\sum_{w \in \bar{V}} \bar{P}_{vw} < 1$. Then $\bar{P}$ is
irreducible (in the sense of Markov chains, i.e., for each \( v, w \in \bar{V} \) there exists \( t \) to that \( \bar{P}^t_{vw} \neq 0 \) and the spectral radius of \( \bar{P} \) is strictly less than 1 (see Corollary 6.2.28 in Horn and Johnson (1986), for instance), so that the Neumann series \( \sum_{t=0}^{\infty} \bar{P}^t \) converges. The Neumann series expansion for \( \overline{L}^{-1} \) yields \( \overline{L}^{-1} = \sum_{t=0}^{\infty} (I - \bar{D}^{-1} \bar{L})^t \bar{D}^{-1} = \sum_{t=0}^{\infty} \bar{P}^t \bar{D}^{-1} \), or, entry-wise, \( \overline{L}^{-1}_{vw} = \frac{1}{\bar{d}_v} \sum_{t=0}^{\infty} \bar{P}^t_{vw} \). As \( \bar{P}^t_{vw} = P_v(X_t = w, T_2 > t) \) by Proposition 15, by the Monotone convergence theorem we can take the summation inside the expectation and get

\[
\sum_{t=0}^{\infty} \bar{P}^t_{vw} = \sum_{t=0}^{\infty} E_v[1_{X_t = w} 1_{T_2 > t}] = E_v \left[ \sum_{t=0}^{T_2} 1_{X_t = w} \right],
\]

where in the last step we used that \( X_{T_2} = \bar{z} \) and \( \bar{z} \neq w \). Recall that if \( S \) is a stopping time for the Markov chain \( X := X_0, X_1, X_2, \ldots \) then by the strong Markov property we know that, conditionally on \( \{ S < \infty \} \) and \( \{ X_S = w \} \), the chain \( X_S, X_{S+1}, X_{S+2}, \ldots \) has the same law as a time-homogeneous Markov chain \( Y := Y_0, Y_1, Y_2, \ldots \) with transition matrix \( P \) and initial condition \( Y_0 = w \), and \( Y \) is independent of \( X_0, \ldots, X_S \). The hitting times \( T_w \) and \( T_2 \) are two stopping times for \( X \), and so is their minimum \( S := \min\{T_w, T_2\} \). As \( X_S = w \) or \( X_S = \bar{z} \), we have

\[
E_v \left[ \sum_{t=0}^{T_2} 1_{X_t = w} \right] = E_v \left[ \sum_{t=0}^{T_2} 1_{X_t = w} \bigg| X_S = w \right] P_v(X_S = w),
\]

where we used that, conditionally on \( \{ X_S = \bar{z} \} = \{ T_w > T_2 \} \), clearly \( \sum_{t=0}^{T_2} 1_{X_t = w} = 0 \). Conditionally on \( \{ X_S = w \} = \{ T_w < T_2 \} = \{ S = T_w \} \), we have \( T_2 = S + \inf\{t \geq 0 : X_{S+t} = \bar{z}\} \), and the strong Markov property yields (note that the event \( \{ S < \infty \} \)) has probability one from any starting point, as the graph \( G \) is connected and the chain will almost surely eventually hit either \( w \) or \( \bar{z} \) that \( E_v[\sum_{t=0}^{T_2} 1_{X_t = w} | X_S = w] \) can be written as \( E_v[\sum_{t=0}^{\inf\{t\geq0|X_{S+t} = \bar{z}\}} 1_{X_{S+t} = w} | X_S = w] = E_v[\sum_{t=0}^{T_2} 1_{X_t = w} | X_S = w] \). Putting everything together we have \( \overline{L}^{-1}_{vw} = \frac{1}{\bar{d}_v} E_w[\sum_{t=0}^{T_2} 1_{X_t = w}] P_v(T_w < T_2) \). As \( P_w(T_w < T_2) = 1 \), clearly \( \overline{L}^{-1}_{vw} = \frac{1}{\bar{d}_v} E_w[\sum_{t=0}^{T_2} 1_{X_t = w}] \) so that \( \overline{L}^{-1}_{vw} = \overline{L}^{-1}_{ww} P_v(T_w < T_2) \). The argument just presented extends easily to the case when \( G \) is not connected. In fact, in this case the matrix \( \bar{P} \) has a block structure, where each block corresponds to a connected component and to a sub-stochastic submatrix, so that the argument above can be applied to each block separately. \( \blacksquare \)

The following result relates the inverse of the reduced Laplacian \( \overline{L} \) with the pseudoinverse of the Laplacian \( L \), which we denote by \( L^+ \). It is proved in Fouss et al. (2007)[Appendix B].

**Proposition 17** For \( v, w \in \bar{V} \), \( \overline{L}^{-1}_{vw} = (e_v - e_{\bar{z}})^T L^+ (e_w - e_{\bar{z}}) \).

Proposition 16 and Proposition 17 allow us to relate the quantity \( L^+ \) to the difference of the Green’s function of the random walk \( X \), as we discuss next.

**C.2 Laplacians and random walks**

We now relate the Moore-Penrose pseudoinverse of the Laplacian \( L := D - W \) with the Green’s function \( (u, v) \in V \times V \to \sum_{t=0}^{\infty} P^t_{uv} = E_u[\sum_{t=0}^{\infty} 1_{X_t = v}] \) of the random walk, which
represents the expected number of times the Markov chain $X$ visits site $v$ when it starts from site $u$. Notice that as the graph $G$ is finite and connected, then the Markov chain $X$ is recurrent and the Green’s function itself equals infinity for any $u, v \in V$. In fact, the following result involves differences of the Green’s function, not the Green’s function itself.

**Lemma 18** For any $u, v, w, z \in V$, we have

\[
(e_u - e_v)^T L^+(e_w - e_z) = \sum_{t=0}^{\infty} (e_u - e_v)^T P^t \left( \frac{e_w}{d_w} - \frac{e_z}{d_z} \right).
\]

**Proof** Using first Proposition 17 and then Proposition 16 we obtain, for any $u, v, w, z \in V$ (choose $\bar{z}$ to be $z$ in Section C.1),

\[
(e_u - e_v)^T L^+(e_w - e_z) = (e_u - e_z)^T L^+(e_w - e_z) - (e_v - e_z)^T L^+(e_w - e_z) = \mathcal{L}_{uw}^{-1} - \mathcal{L}_{vw}^{-1} = (e_u - e_z)^T L^+(e_w - e_z) \{\mathbf{P}_u(T_w < T_z) - \mathbf{P}_v(T_w < T_z)\}.
\]

From (3.27) in the proof of Proposition 3.10 in Chapter 3 in Aldous and Fill (2002), upon identifying $v \rightarrow u, x \rightarrow v, v_0 \rightarrow w, a \rightarrow z$, we immediately have the following relation between the difference of potentials and hitting times of the random walk $X$: $\mathbf{P}_u(T_w < T_z) - \mathbf{P}_v(T_w < T_z) = \pi_w \mathbf{P}_w(T_z < T_w) \{\mathbf{E}_u T_z - \mathbf{E}_u T_w + \mathbf{E}_v T_w - \mathbf{E}_v T_z\}$, where $\pi_v := \frac{d_v}{\sum_{v \in V} d_v}$ is the $v$-th component of the stationary distribution of the random walk $X$, and $T^+_v := \inf\{t \geq 1 : X_t = v\}$. From Corollary 8 in Chapter 2 in Aldous and Fill (2002), we have

\[
\pi_w \mathbf{P}_w(T_z < T_w^+) = \begin{cases} \mathbf{E}_w T_z + \mathbf{E}_z T_w & \text{if } w \neq z, \\ \frac{1}{\pi_w} \sum_{t=0}^{\infty} (P^t_{uw} - P^t_{vw}) & \text{if } w = z, \end{cases}
\]

and we recall the connection between commute times and effective resistance (see, e.g., Corollary 3.11 in Aldous and Fill (2002)): $\mathbf{E}_w T_z + \mathbf{E}_z T_w = (e_u - e_z)^T L^+(e_w - e_z) \sum_{v \in V} d_v$.

Lemma 3.3 in Friedrich and Sauerwald (2010) gives $\mathbf{E}_u T_z - \mathbf{E}_u T_w = \frac{1}{\pi_w} \sum_{t=0}^{\infty} (P^t_{uw} - P^t_{vw})$, and $\mathbf{E}_w T_w - \mathbf{E}_v T_w = \frac{1}{\pi_w} \sum_{t=0}^{\infty} (P^t_{vw} - P^t_{uw})$. The proof of the lemma follows by combining everything together. 

We have the following corollary of Lemma 18, which immediately yields the proof of Theorem 7 in Section 4.

**Corollary 19** For $u, v \in V$, $f = (f_z)_{z \in V} \in \mathbb{R}^V$ with $\mathbf{1}^T f = 0$,

\[
(e_u - e_v)^T L^+ f = \sum_{z \in V} \sum_{t=0}^{\infty} (P^t_{uz} - P^t_{vz}) \frac{f_z}{d_z}.
\]

**Proof** From Lemma 18, by summing the quantity $(e_u - e_v)^T L^+(e_w - e_z)$ over $z \in V$, recalling that $\sum_{z \in V} e_z = \mathbf{1}$ and $L^+ \mathbf{1} = 0$ we have $(e_u - e_v)^T L^+ e_w = \sum_{t=0}^{\infty} (P^t_{uw} - P^t_{vw}) \frac{1}{d_w} - \frac{1}{|V|} \sum_{z \in V} \sum_{t=0}^{\infty} (P^t_{uz} - P^t_{vz}) \frac{1}{d_z}$. The identity in the statement of the Lemma follows easily as $f = \sum_{w \in V} f_w e_w$ and $\sum_{w \in V} f_w = 0$ by assumption.
Appendix D. Decay of correlation

This appendix is devoted to the proofs of the decay of correlation properties stated in Section 4, namely, Theorem 8 (set-to-point) and Lemma 10 (point-to-set). These proofs rely on the sensitivity analysis for the network flow problem established in Theorem 7.

Henceforth, consider the general setting introduced in Appendix C, and let \( d \) denote the graph-theoretical distance on \( G \) : that is, \( d(u, v) \) denotes the length of the shortest path between vertex \( u \) and vertex \( v \). Note that \( d(u, v) = \inf \{ t \geq 0 : P^t_{uv} \neq 0 \} \), as we assumed that to each edge \( \{v, w\} \in E \) is associated a non-negative weight \( W_{vw} = W_{uw} > 0 \). Let \( n := |V| \), and let \(-1 \leq \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 < \lambda_1 = 1 \) be the eigenvalues of \( P \). Define \( \lambda := \max \{|\lambda_2|, |\lambda_n|\} \).

The backbone behind the proof of Theorem 8 and Lemma 10 is given by the following lemma.

Lemma 20 For any \( U, Z \subseteq V \) and \((f_z)_{z \in Z} \in \mathbb{R}^Z \) we have
\[
\sqrt{\frac{1}{2} \sum_{u,v \in U : \{u,v\} \in E} \left( \sum_{z \in Z} \sum_{t=0}^\infty (P^t_{uz} - P^t_{vz}) f_z \right)^2} \leq \alpha \frac{\lambda_{d(U,Z)}}{1 - \lambda} \sqrt{\sum_{z \in Z} f_z^2 d_z},
\]
with \( \alpha := \frac{\max_{u \in U} \sqrt{2N(u) - U}}{\min_{v \in U} \sqrt{d_v}} \).

Proof Let \( \Gamma := D^{1/2} P D^{-1/2} = D^{-1/2} W D^{-1/2} \). This matrix is clearly similar to \( P \) and symmetric. Let denote by \( \psi_1, \ldots, \psi_1 \) the orthonormal eigenvectors of \( \Gamma \) corresponding to the eigenvalues \( \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 \leq \lambda_1 \), respectively. By substitution, it is easy to check that \( \sqrt{\pi} \equiv (\sqrt{\pi_v})_{v \in V} \) is an eigenvector of \( \Gamma \) with eigenvalue equal to 1, where we recall that \( \pi_v = d_v / \sum_{v \in V} d_v \). Since this eigenvector has positive entries, it follows by the Perron-Frobenius theory that \(-1 \leq \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_2 < \lambda_1 = 1 \) and that \( \psi_1 = \sqrt{\pi} \). As \( \Gamma = \sum_{k=1}^n \lambda_k \psi_k \psi_k^T \), by the orthonormality of the eigenvectors we have, for \( t \geq 0, u, z \in V \),
\[
P^t_{uz} = (D^{-1/2} \Gamma^t D^{1/2})_{uz} = \pi_z + \sum_{k=2}^n \lambda_k \psi_{ku} \psi_{kz} \sqrt{\frac{d_z}{d_u}},
\]
where \( \psi_{ku} \equiv (\psi_k)_u \) is the \( u \)-th component of \( \psi_k \). As \( P^t_{uz} = 0 \) if \( d(u, z) > t \), we have \( P^t_{uz} = P^t_{vz} = 1_{d(U,Z) \leq t} (P^t_{uz} - P^t_{vz}) \) for any \( u, v \in U, z \in Z \). Hence, for any \( u, v \in U \), let \( g_{uv} := \sum_{z \in Z} \sum_{t=0}^\infty (P^t_{uz} - P^t_{vz}) f_z \), which equals
\[
g_{uv} = \sum_{k=2}^n \left( \frac{\psi_{ku}}{\sqrt{d_u}} - \frac{\psi_{kv}}{\sqrt{d_v}} \right) \sum_{z \in Z} \psi_{kz} \sqrt{d_z} f_z \sum_{t=d(U,Z)}^\infty \lambda_k^t.
\]
As \( \lambda < 1 \) by assumption, the geometric series converges for any \( k \neq 1 \). If we define \( h_u := \sum_{k=2}^n \frac{\lambda_k^{d(U,Z)}}{1 - \lambda_k} \frac{\psi_{ku}}{\sqrt{d_u}} \sum_{z \in Z} \psi_{kz} \sqrt{d_z} f_z \) for each \( u \in V \), we have \( g_{uv} = h_u - h_v \), and the triangle inequality for the \( \ell_2 \)-norm yields \( (\sum_{u,v \in U : \{u,v\} \in E} g_{uv}^2)^{1/2} \leq 2 (\sum_{u,v \in U : \{u,v\} \in E} h_u^2)^{1/2} \) which is upper-bounded by \( 2 \sqrt{\max_{u \in U} [N(u) \cap U] \left( \sum_{u \in U} h_u^2 \right)^{1/2}} \), where the factor 2 comes
by the symmetry between \( u \) and \( v \). Expanding the squares and using that \( |\lambda_k| \leq \lambda \) for each \( k \neq 1 \), we get
\[
d_u h_u^2 \leq \frac{\lambda^{2d(U, Z)}}{(1 - \lambda)^2} \sum_{k=1}^{n} \psi_{ku}^2 \left( \sum_{z \in Z} \psi_{kz}^2 d_z f_z^2 + \sum_{z \neq z'} \psi_{kz} \psi_{kz'} \sqrt{d_z d_{z'} f_z f_{z'}} \right) + \sum_{k \neq k' \geq 2} \frac{\lambda^{d(U, Z)} \lambda^{d(U, Z)}}{(1 - \lambda_k)(1 - \lambda_{k'})} \psi_{ku} \psi_{k'u} \sum_{z, z'} \psi_{kz} \psi_{kz'} \sqrt{d_z d_{z'} f_z f_{z'}} ,
\]
where we also used that \( \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} x_k \) if \( x_1, \ldots, x_n \) are non-negative numbers. Let \( \Psi \) denote the matrix having the eigenvectors \( \psi_1, \ldots, \psi_n \) in its columns, namely, \( \Psi_{uk} = (\psi_k)_u = \psi_{ku} \). This is an orthonormal matrix, so both its columns and rows are orthonormal, namely, \( \sum_{u=1}^{n} \psi_{ku} \psi_{k'u} = 1_{k=k'} \) and \( \sum_{k=1}^{n} \psi_{ku} \psi_{kv} = 1_{u=v} \). Using this fact,
\[
\sum_{u \in U} h_u^2 \leq \frac{\sum_{u \in V} d_u h_u^2}{\min_{u \in U} d_u} \leq \frac{1}{\min_{u \in U} d_u} \frac{\lambda^{2d(U, Z)}}{(1 - \lambda)^2} \sum_{z \in Z} d_z f_z^2 ,
\]
and the proof follows by putting all the pieces together, realizing that the upper bound in the statement of the lemma corresponds to \( \frac{1}{\sqrt{2}} (\sum_{u, v \in U} \Psi_{uv} g_{uv})^{1/2} \).

We are finally ready to present the proof of Theorem 8.

**Proof** [Proof of Theorem 8] Consider the setting developed in Section 4. Fix \( \varepsilon \in \mathbb{R} \). From Theorem 7 we have \( \frac{dx^*(b(\varepsilon))}{d\varepsilon} = \Sigma(b(\varepsilon)) A^T L(b(\varepsilon)) + \frac{\partial b(\varepsilon)}{d\varepsilon} \) or, entry-wise, for any \( (u, v) \in E \),
\[
\frac{dx^*(b(\varepsilon))[u,v]}{d\varepsilon} = W(b(\varepsilon))_{uv}(e_u - e_v)^T L(b(\varepsilon)) + \frac{\partial b(\varepsilon)}{d\varepsilon} .
\]
Define \( U := V_{\bar{F}} \), and let \( (U, \bar{F}) \) be the undirected graph associated to \( (U, F) \) (see Remark 1). Clearly, \( (\sum_{\varepsilon \in \mathbb{E}} (\frac{dx^*(b(\varepsilon))}{d\varepsilon})^2)^{1/2} \) is upper-bounded by
\[
\gamma(b(\varepsilon)) \sqrt{\frac{1}{2} \sum_{u,v \in U : \{u,v\} \in E} \left( (e_u - e_v)^T L(b(\varepsilon)) + \frac{\partial b(\varepsilon)}{d\varepsilon} \right)^2} ,
\]
where \( \gamma(b) := \max_{u,v \in U} W(b)_{uv} \) for any \( b \in \text{Im}(A) \). Corollary 19 yields (choosing \( f = \frac{\partial b(\varepsilon)}{d\varepsilon} \)), using that \( 1^T f = 0 \) and \( f_e \neq 0 \) if and only if \( v \in Z \) that \( (e_u - e_v)^T L(b(\varepsilon)) + \frac{\partial b(\varepsilon)}{d\varepsilon} \) equals \( \sum_{z \in Z} \sum_{t=0}^{\infty} (P(b(\varepsilon))_{uz}^t - P(b(\varepsilon))_{uz}) \frac{\partial b(\varepsilon)}{d\varepsilon}^z \), so that (5) reads
\[
\gamma(b(\varepsilon)) \sqrt{\frac{1}{2} \sum_{u,v \in U : \{u,v\} \in E} \left( \sum_{z \in Z} \sum_{t=0}^{\infty} (P(b(\varepsilon))_{uz}^t - P(b(\varepsilon))_{uz}) \frac{\partial b(\varepsilon)}{d\varepsilon}^z \right)^2} .
\]
Lemma 20 yields (choosing \( f_z = \frac{1}{d\varepsilon} \frac{\partial b(\varepsilon)}{d\varepsilon}^z \)) that the previous quantity is upper-bounded by
\[
\gamma(b(\varepsilon)) \alpha(b(\varepsilon)) \frac{\lambda(b(\varepsilon))^d(U, Z)}{1 - \lambda(b(\varepsilon))} \sqrt{\sum_{z \in Z} \left( \frac{\partial b(\varepsilon)}{d\varepsilon}^z \right)^2} \frac{1}{\lambda(b(\varepsilon))} .
\]
with \( \alpha(b) := \frac{\max_{v \in U} \sqrt{2|N(v)\cap U|}}{\min_{u \in U} \sqrt{d(b)_v}} \) for any \( b \in \text{Im}(A) \). Combining everything together, we obtain
\[
\left\| \frac{dx^*(b(\varepsilon))}{d\varepsilon} \right\|_F \leq c(b(\varepsilon)) \frac{\lambda(b(\varepsilon)) d(U,Z)}{1 - \lambda(b(\varepsilon))} \left\| \frac{db(\varepsilon)}{d\varepsilon} \right\|_Z,
\]
where \( c(b) := \frac{\max_{v \in U} \sqrt{2|N(v)\cap U|}}{\min_{u \in U} \sqrt{d(b)_v}} \gamma(b) \) for \( b \in \text{Im}(A) \). The proof follows by taking the supremum over \( b \in \text{Im}(A) \). \( \blacksquare \)

The proof of Lemma 10 follows analogously from the proof of Theorem 8, upon exploiting the symmetry of the pseudoinverse of the graph Laplacian. \( \blacksquare \)

**Proof** [Proof of Lemma 10] Consider the setting developed in Section 4. From Theorem 7 we immediately have, for \( e = (u, v) \in \overline{E} \) and \( b \in \mathbb{R}^V \) such that \( 1^T b = 0, \nabla_e x^*(b) = \Sigma(b) A^T L(b)^+(e_u - e_v) \) or, entry-wise, for any \( f = (w, z) \in \overline{E}, \nabla_e x^*(b)_f = W(b)_w z (e_u - e_v)^T L(b)^+(e_w - e_z) \), where we used the fact that the matrix \( L(b)^+ \) is symmetric. Let \( U := V_F \). Clearly, \( (\Sigma_{e \in E}(\nabla_e x^*(b)_f)^2)/2 \) is upper-bounded by \( W(b)_w z (\frac{1}{2} \Sigma_{u,v \in U; \{u,v\} \in E} ((e_u - e_v)^T L(b)^+(e_w - e_z))^2) \). From this point onward, the proof mirrors exactly the proof of Theorem 8, from (5) onward, upon substituting \( \max_{u,v \in U} W(b(\varepsilon))_{uv} \rightarrow W(b)_{uv}, L(b(\varepsilon)) \rightarrow L(b), \frac{b(\varepsilon)}{d\varepsilon} \rightarrow e_w - e_z \), and upon choosing \( Z = V_{\{f\}} = \{w, z\} \). \( \blacksquare \)

**Appendix E. Localized algorithm**

This section is devoted to the proof of Theorem 12, which states error bounds for the localized projected gradient descent algorithm. The proof relies on the decay of correlation property established in Theorem 8 for the network flow problem. Recall that the constants appearing in the bounds in Theorem 12 do not depend on the choice of the subgraph \( \tilde{G}' \) of \( \tilde{G} \), but depend only on \( \mu, Q, k_+, \) and \( k_- \). To prove this type of bounds, we first need to develop estimates to relate the eigenvalues of weighted subgraphs to the eigenvalues of the corresponding unweighted graph.

**E.1 Eigenvalues interlacing**

Let \( G = (V, E) \) be a simple (i.e., no self-loops, and no multiple edges), connected, undirected graph, with vertex set \( V \) and edge set \( E \). Let \( B \in \mathbb{R}^{V \times V} \) be the vertex-vertex adjacency matrix of the graph, which is the symmetric matrix defined as \( B_{uv} := 1 \) if \( \{u, v\} \in E, B_{uv} := 0 \) otherwise. If \( n := |V|, \) denote by \( \mu_n \leq \mu_{n-1} \leq \cdots \leq \mu_2 \leq \mu_1 \) the eigenvalues of \( B \). Let \( G' = (V', E') \) be a connected subgraph of \( G \). Assume that to each edge \( \{u, v\} \in E' \) is associated a non-negative weight \( W_{uv} = W_{vu} > 0 \), and let \( W_{uv} = 0 \) if \( \{u, v\} \notin E \). Let \( D' \) be a diagonal matrix with entries \( D'_{vv} = \sum_{w \in V} W'_{vw} \) for each \( v \in V' \). Let \( P' := D'^{-1} W' \). If \( m := |V'|, \) denote by \( \lambda'_m \leq \lambda'_{m-1} \leq \cdots \leq \lambda'_2 \leq \lambda'_1 \) the eigenvalues of \( P' \). The following proposition relates the eigenvalues of \( P' \) to the eigenvalues of \( B \). In particular, we provide a bound for the second largest eigenvalue in magnitude of \( P' \) with respect to the second largest eigenvalue in magnitude of \( B \), uniformly over the choice of \( G' \).
Proposition 21 (Eigenvalues interlacing) Let \( w_- \leq W_{vw} \leq w_+ \) for any \( \{v, w\} \in E \), for some constants \( w_- , w_+ > 0 \). Let \( k_- \) and \( k_+ \) be, respectively, the min and max degree of \( G \). Then,
\[
1 - \frac{w_+ k_+}{w_- k_-} + \frac{w_+}{w_- k_-} \mu_{i+n-m} \leq \lambda'_i \leq 1 - \frac{w_- k_-}{w_+ k_+} + \frac{w_-}{w_+ k_+} \mu_i.
\]

Therefore, if \( \lambda' := \max\{\lambda'_2, |\lambda'_m|\} \) and \( \mu := \max\{\mu_2, |\mu_n|\} \), we have \( \lambda' \leq \frac{w_+ k_+}{w_- k_-} - 1 + \frac{w_-}{w_+ k_+} \mu \).

**Proof** Consider \( \Gamma' := D'^{1/2}P'D'^{-1/2} = D'^{-1/2}W'D'^{-1/2} \). As this matrix is similar to \( P' \), it shares the same eigenvalues with \( P' \). Let \( L' := D' - W' \) be the Laplacian associated to \( G' \). The Courant-Fischer Theorem yields
\[
\lambda'_i = \max_{S \subseteq \mathbb{R}^m \atop \dim(S) = i} \min_{x \in S} \frac{x^T \Gamma' x}{x^T x} = 1 + \max_{S \subseteq \mathbb{R}^m \atop \dim(S) = i} \min_{y \in S} \frac{-y^T L'y}{y^T D'y},
\]
where we used that \( x^T \Gamma' x = x^T x - y^T L'y \) with \( y := D'^{-1/2}x \), and that the change of variables \( y = D'^{-1/2}x \) is non-singular (note that as \( G' \) is connected, then \( D' \) has non-zero entries on the diagonal). The Lagrange quadratic form yields \( y^T L'y = \frac{1}{2} \sum_{u,v \in V'} W_{uv}(y_u - y_v)^2 \), which is upper-bounded by \( w_+ + \frac{1}{2} \sum_{u,v \in V'} B_{uv}(y_u - y_v)^2 = w_+ y^T \mathcal{L}' y \), where \( \mathcal{L}' \) is the Laplacian of the unweighted graph \( G' = (V', E') \equiv (V', E', B') \) with \( B' := B_{v' w'} \). Note that we have \( \mathcal{L}' = \mathcal{K}' - B' \), where \( \mathcal{K}' \) is diagonal and \( \mathcal{K}'_{v w} = \sum_{w' \in V'} B'_{w' w} \) is the degree of vertex \( v \in V' \) in \( G' \). As \( y^T \mathcal{K}' y = \sum_{v \in V'} \mathcal{K}_{v v} y_v^2 \leq k_+ y^T y \), we have \( y^T L'y \leq w_+ k_+ y^T y - w_+ y^T B'y \). At the same time, \( y^T D'y \geq w_- k_- y^T y \). Therefore, \( \lambda'_i \geq 1 - \frac{w_+ k_+}{w_- k_-} + \frac{w_-}{w_+ k_+} \max_{S \subseteq \mathbb{R}^m \atop \dim(S) = i} \min_{y \in S} \frac{-y^T B'y}{y^T y} \), and by the Courant-Fischer Theorem the right-hand side equals \( 1 - \frac{w_+ k_+}{w_- k_-} + \frac{w_-}{w_+ k_+} \mu_i \), where \( \mu'_i \leq \mu'_{i+1} \leq \cdots \leq \mu'_{2} \leq \mu'_1 \) are the eigenvalues of \( B' \). Analogously, it is easy to prove that \( \lambda'_i \leq 1 - \frac{w_- k_-}{w_+ k_+} + \frac{w_+}{w_- k_-} \mu'_i \). As \( B' \) is a principal submatrix of \( B \), the eigenvalue interlacing theorem for symmetric matrices yields \( \mu_{i+n-m} \leq \mu'_i \leq \mu_i \), and we have \( \lambda'_i \leq 1 - \frac{w_- k_-}{w_+ k_+} + \frac{w_+}{w_- k_-} \mu_i \), where \( \alpha := 1 - \frac{w_- k_-}{w_+ k_+} + \frac{w_+}{w_- k_-} \mu_i \), and \( \beta := 1 - \frac{w_- k_-}{w_+ k_+} + \frac{w_+}{w_- k_-} \mu_i \).

E.2 Proof of Theorem 12

We now present the proof of Theorem 12. The proof relies on repeatedly applying Theorem 8 in Section 4 (which captures the decay of correlation for the network flow problem) and the fundamental theorem of calculus.

**Proof** [Proof of Theorem 12] Consider the setting of Section 5.

**Bias term.** Let us first bound the bias outside \( \widehat{E}' \). Let \( n := |V'| \), and for each \( b \in \text{Im}(A) \) let \( -1 \leq \lambda_n(b) \leq \lambda_{n-1}(b) \leq \cdots \leq \lambda_2(b) < \lambda_1(b) = 1 \) be the eigenvalues of \( P(b) \). Let \( \lambda(b) := \max\{\lambda_2(b), |\lambda_n(b)|\} \) and \( \lambda := \sup_{b \in \text{Im}(A)} \lambda(b) \). Define \( b(\varepsilon) := b + \varepsilon p \), for any
non-negative real number $\varepsilon \geq 0$. If $e \in \tilde{E}^C$, then $T'_{b(\varepsilon)}(x^*(b))_e = x^*(b)_e$ and

$$
\text{Bias}(p, \tilde{G}')_e = x^*(b(1))_e - x^*(b(0))_e = \int_0^1 d\varepsilon \frac{dx^* (b(\varepsilon))_e}{d\varepsilon}.
$$

By the triangle inequality for the $\ell_2$-norm and Theorem 8,

$$
\| \text{Bias}(p, \tilde{G}') \|_{E^C} \leq \sup_{\varepsilon \in \mathbb{R}} \left\| \frac{dx^* (b(\varepsilon))}{d\varepsilon} \right\|_{E^C} \leq c\|p\| \frac{\lambda(d(\Delta(\tilde{G}), Z))}{1 - \lambda},
$$

where we used that $\sup_{\varepsilon \in \mathbb{R}} \left\| \frac{db(\varepsilon)}{d\varepsilon} \right\|_{Z} = \|p\|$, as $\frac{db(\varepsilon)}{d\varepsilon} = p_v$ for $v \in Z$ and $\frac{db(\varepsilon)}{d\varepsilon} = 0$ for $v \notin Z$, and $c := \sqrt{2k_+Q/k_-}$.

Let us now consider the bias inside $\tilde{E}'$. Let $A' := A_{V', \tilde{E}'} \in \mathbb{R}^{V' \times \tilde{E}'}$ be the vertex-edge adjacency matrix of the subgraph $\tilde{G}'$. For $b' \in \text{Im}(A') \subseteq \mathbb{R}^{V'}$, consider the following optimization problem over $x' \in \mathbb{R}^{	ilde{E}'}$:

$$
\text{minimize} \quad f'(x') := \sum_{e \in \tilde{E}'} f_e(x'_e)
$$

subject to \( A' x' = b' \),

and denote its unique optimal point as $x'^* (b') := \arg \min \left\{ f'(x') : x' \in \mathbb{R}^{	ilde{E}'} , A' x' = b' \right\} \in \mathbb{R}^{	ilde{E}'}$. For any $\varepsilon > 0$, $\theta > 0$, define

$$
b'(\varepsilon, \theta) := b(\varepsilon)_{V'} - A_{V', \tilde{E}^C} x^*(b(\theta))_{\tilde{E}^C}. \quad (6)
$$

Without loss of generality, we can index the elements of $V'$ and $\tilde{E}'$ so that the matrix $A$ has the following structure:

$$
A = \begin{pmatrix}
A_{V', \tilde{E}'} & A_{V', \tilde{E}^C} \\
A_{V^C \setminus \tilde{E}'} & A_{V^C \setminus \tilde{E}^C}
\end{pmatrix} = \begin{pmatrix}
A' & A_{V', \tilde{E}^C} \\
0 & A_{V^C \setminus \tilde{E}^C}
\end{pmatrix}.
$$

For any $x$ that satisfies the flow constraints on $\tilde{E}^C$ with respect to $b(\varepsilon)$, i.e., $A_{V \setminus V', \tilde{E}^C} x_{\tilde{E}^C} = b(\varepsilon)_{V \setminus V'}$,

$$
\lim_{t \to \infty} T_{b+p}^t(x)_{\tilde{E}'} = x'^* (b(1)_{V'} - A_{V', \tilde{E}^C} x_{\tilde{E}^C}).
$$

Clearly $x^*(b)$ satisfies the flow constraints on $\tilde{E}^C$ with respect to $b(1)$, as $p$ is supported on $V'$ so that $b(\varepsilon)_{V^C} = b_{V^C}$. Recalling the definition of $b'(\varepsilon, \theta)$ in (6), we then have \( (\lim_{t \to \infty} T_{b+p}^t(x^*(b(\varepsilon)))_{\tilde{E}'} = x'^* (b'(1, 0)) \). On the other hand, as $x^*(b(1))$ is clearly a fixed point of the map $T_{b(1)}'$, we can characterize the components of $x^*(b(1))$ supported on $\tilde{E}'$ as

$$
x^*(b(1))_{\tilde{E}'} = \left( \lim_{t \to \infty} T_{b(1)}^t(x^*(b(1))) \right)_{\tilde{E}'} = x'^* (b'(1, 1)).
$$
It is easy to check that $b'(\varepsilon, \theta) \in \text{Im}(A')$ for each value of $\varepsilon$ and $\theta$. In fact, as $\tilde{G}'$ is connected by assumption, then $\text{Im}(A')$ corresponds to the subspace of $\mathbb{R}^{V'}$ orthogonal to the all-ones vector $\mathbf{1}$. We have $1^T b'(\varepsilon, \theta) = 1^T b_{V'} + \varepsilon 1^T p_{V'} - 1^T A_{V', \tilde{E}C} x^*(b(\theta))_{\tilde{E}C}$. Note that $1^T p_{V'} = 0$ by assumption. Also, $0 = 1^T b = 1^T b_{V'} + 1^T b_{V'C}$ (note the different dimension of the all-ones vectors) so that $1^T b_{V'} = -1^T b_{V'C}$. Analogously, as $1^T A = 0^T$, we have $1^T A_{V', \tilde{E}C} = -1^T A_{V'C, \tilde{E}C}$. Hence,

$$1^T b'(\varepsilon, \theta) = -1^T b_{V'C} + 1^T A_{V'C, \tilde{E}C} x^*(b(\theta))_{\tilde{E}C} = 0^T,$$

where the last equality follows as clearly $A_{V'C, \tilde{E}C} x^*(b(\theta))_{\tilde{E}C} = b_{V'C}$. Therefore, for $e \in \tilde{E}'$, 

$$\text{Bias}(p, \tilde{G}')_e = \int_0^1 d\theta \left| \frac{dx^*(b'(1, \theta))}{d\theta} \right|_e.$$

For each $b' \in \text{Im}(A')$, let $W'(b') \in \mathbb{R}^{V' \times V'}$ be a symmetric matrix defined as $W'(b')_{uv} = (\partial^2 f_e(x^*(b'))/\partial x^2)_{uv}^{-1}$ if either $e = (u, v) \in \tilde{E}$ or $e = (v, u) \in \tilde{E}$, and $W'(b')_{uv} := 0$ otherwise. Let $D'(b') \in \mathbb{R}^{V' \times V'}$ be a diagonal matrix with entries $D'(b')_{vv} = \sum_{u \in V'} W'(b')_{uv}$. Let $P'(b') := D'(b')^{-1} W'(b')$. If $m := |V|$, let $\lambda_m(b') \leq \lambda_{m-1}(b') \leq \cdots \leq \lambda_1(b') = 1$ be the eigenvalues of $P'(b')$ (where this characterization holds as $G'$ is connected by assumption). Define $\lambda'(b') := \max\{|\lambda_2(b')|, |\lambda_3(b')|\}$ and $\lambda' := \sup_{b' \in \text{Im}(A')} \lambda'(b')$. Proceeding as above, applying Theorem 8 to the optimization problem defined on $\tilde{G}'$ (recall that $G'$ is connected by assumption), we get

$$\|\text{Bias}(p, \tilde{G}')\|_{\tilde{E}} \leq \sup_{\theta \in \mathbb{R}} \frac{\|dx^*(b'(1, \theta))\|_{\tilde{E}}}{\|\partial \lambda' / \partial \theta\|_{\Delta(\tilde{G})}},$$

which is upper-bounded by $c \frac{\| \partial \lambda' / \partial \theta \|_{\Delta(\tilde{G})}}{1 - \lambda'} \leq \sup_{\theta \in \mathbb{R}} \frac{\| \partial \lambda' / \partial \theta \|_{\Delta(\tilde{G})}}{1 - \lambda'}$, where we used that $\frac{\partial \lambda' / \partial \theta}{\partial \theta} = 0$ if $v \in V' \setminus \Delta(\tilde{G})$, and clearly $d(V', \Delta(\tilde{G})) = 0$ as $\Delta(\tilde{G}) \subseteq V'$. For $v \in \Delta(\tilde{G})$ we have

$$\frac{\partial \lambda' / \partial \theta}{\partial \theta} = -\sum_{e \in \tilde{E}C} A_{ve} \frac{dx^*(b(\theta))}{d\theta}.$$  

If $\tilde{F}(v) := \{ e \in \tilde{E}C : e = (u, v) \text{ or } e = (v, u) \text{ for some } u \in V'^C \}$ denotes the set of edges that are connected to $v$ but do not belong to $\tilde{E}$, we have $(\partial \lambda'(1, \theta) / \partial \theta)^2 \leq (\sum_{e \in \tilde{F}(v)} |d\lambda'(b(\theta))/d\lambda|)^2$, which by Jensen’s inequality admits $\| \tilde{F}(v) \| \sum_{e \in \tilde{F}(v)} (d\lambda'(b(\theta))/d\theta)^2 \right\|_2$ as a upper bound. As $\max_{v \in \Delta(\tilde{G})} \| \tilde{F}(v) \| \leq k + 1$, applying Theorem 8 as above we get

$$\|\text{Bias}(p, \tilde{G}')\|_{\tilde{E}} \leq c^2 \sqrt{k + 1} \| p \|_{\tilde{E}} \| \lambda'(\Delta(\tilde{G})) \|^2 \frac{\lambda'(\Delta(\tilde{G}))}{1 - \lambda'}.$$  

Therefore, $\|\text{Bias}(p, \tilde{G}')\|_{\tilde{E}} \leq c^2 \sqrt{k + 1} \| p \|_{\tilde{E}} \| \lambda'(\Delta(\tilde{G})) \|^2 \frac{\lambda'(\Delta(\tilde{G}))}{1 - \lambda'}$. By the triangle inequality for the $\ell_2$-norm, $\|\text{Bias}(p, \tilde{G}')\|_{\tilde{E}} \leq \|\text{Bias}(p, \tilde{G}')\|_{\tilde{E}} + \|\text{Bias}(p, \tilde{G}')\|_{\tilde{E}}$, so we obtain

$$\|\text{Bias}(p, \tilde{G}')\|_{\tilde{E}} \leq c \| p \|_{\tilde{E}} \| \lambda'(\Delta(\tilde{G})) \|^2 \frac{\lambda'(\Delta(\tilde{G}))}{1 - \lambda'}.$$  

By Proposition 21 we have $\max\{|\lambda, \lambda'| \leq \frac{Qk}{\alpha} - 1 + \frac{Q}{\alpha} \mu$, and the bound for the bias term follows.

**Variance term.** As $(\lim_{t \to \infty} T_{b'p}^{d_t}(x^*(b)))_{\tilde{E}} = x^*(b'(1, 0))$, we get

$$\|\text{Variance}(p, \tilde{G}', t)\|_{\tilde{E}} = \|x^*(b'(1, 0)) - T_{b'p}^{d_t}(x^*(b))_{\tilde{E}}\|_{\tilde{E}} \leq e^{-\frac{d_{t}}{2}} \| x^*(b'(1, 0)) - x^*(b'(0, 0)) \|_{\tilde{E}}.$$  

33
where in the last inequality we used that $x^*(b)_{E'} = x'^*(b'(0,0))$. For each $e \in \vec{E}'$ we have

$$x'^*(b'(1,0))_e - x'^*(b'(0,0))_e = \int_0^1 d\varepsilon \frac{dx'^*(b'(\varepsilon,0))_e}{d\varepsilon},$$

and using the triangle inequality for the $\ell_2$-norm, applying Theorem 8 to the optimization problem defined on $\vec{G}'$,

$$\| \text{Variance}(p, \vec{G}', t) \|_{\vec{E}'} \leq \sup_{\varepsilon \in \mathbb{R}} \left\| \frac{dx'^*(b'(\varepsilon,0))}{d\varepsilon} \right\|_{\vec{E}'} \leq c\|p\| \frac{1}{1 - \lambda'},$$

where we used that $\frac{\partial b'(\varepsilon,0)}{\partial \varepsilon} = \frac{\partial b(\varepsilon)}{\partial \varepsilon} = p_v$ for $v \in Z$ and $\frac{\partial b(\varepsilon)}{\partial \varepsilon} = 0$ for $v \notin Z$, and that $d(V', Z) = 0$ as $Z \subseteq V'$. Clearly, $\text{Variance}(p, \vec{G}', t)_e = 0$ for $e \in \vec{E}' \setminus C$, as $T'_0(x^*(b))_e = x^*(b)_e$. Hence, $\| \text{Variance}(p, \vec{G}', t) \| = \| \text{Variance}(p, \vec{G}', t) \|_{\vec{E}'}$ and the proof is concluded as $\lambda' \leq \frac{Qk_+}{k_-} - 1 + \frac{Q}{k_-} \mu$ by Proposition 21.