Nonlinear Stochastic Programming — With a Case Study in Continuous Switching

Alois Pichler[*] Asgeir Tomasgard[†]

April 6, 2016

Abstract

The optimal solution, as well as the objective of stochastic programming problems vary with the underlying probability measure. This paper addresses stability with respect to the underlying probability measure and stability of the objective.

The techniques presented are employed to make problems numerically tractable, which are formulated by involving numerous scenarios, or even by involving a continuous probability measure. The results justify clustering techniques, which significantly reduce computation times while guaranteeing a desired approximation quality.

The second part of the paper highlights Newton's method to solve the reduced stochastic recourse problems. The techniques presented exploit the particular structure of the recourse function of the stochastic optimization problem. The tools are finally demonstrated on a benchmark problem, which is taken from electrical power flows.

Keywords: Stochastic Optimization, nonlinear programming, risk measures, robust optimization, Wasserstein metrics

Classification: 62H30, 62C12, 90C15

1 Introduction

Stochastic programming problems are often formulated as linear programs. A main reason for that is perhaps that linear programming historically emerged from stochastic optimization: the initial problem Dantzig considers in [7] to develop the linear theory is indeed a stochastic optimization problem. Moreover efficient solvers are available, which can be employed to solve even large scale linear problems. For these reasons genuinely nonlinear problems are often linearized in order to solve the approximating, linear problem by using well established and accepted linear solution methods and techniques. However, important problems are known for which the linear approximation is irrelevant to the understanding of the initial problem (a prominent example in physics, which is often referred to in this context, is the explanation and description of a ship's bow wave).

This paper addresses general *two-stage stochastic optimization* problems involving a general, nonlinear objective, with nonlinear constraints on a general probability space (cf. Shapiro et al. [31] for stochastic optimization). It investigates continuity with respect to the underlying probability

[*] Norwegian University of Science and Technology. Contact: alois.pichler@univie.ac.at

The author gratefully acknowledges support of the Research Council of Norway, grant 207690/E20.

[†] Norwegian University of Science and Technology, Director of *CenSES, Centre for Sustainable Energy Studies*

measure and outlines implications of the particular structure of the problem on numerical solution techniques. A case study, taken from electrical engineering, is presented, for which it was realized recently that the simple, linearized problem does not lead to useful results. The problem in the case study thus has to be solved by accepting its nonlinear complexity.

The general two-stage stochastic optimization problem is

$$\min_{y\in Y} \mathcal{R}\left(\min_{z\in Z(y)} c\,(y,\xi,z)\right), \tag{1}$$

where c is a (cost) function and $z \in Z$ is the *wait-and-see* decision. The inner problem, $\min_{z\in Z(y)} c\,(y,\xi,z)$, depends on the *here-and-now* decision $y \in Y$ and the random variable ξ and thus is random itself. The convex risk measure $\mathcal{R}$ summarizes the different outcomes of the random inner problem in a single real number. In the simplest case, $\mathcal{R}(\cdot) = \mathbb{E}(\cdot)$ is the expectation so that problem (1) reads

$$\min_{y\in Y} \mathbb{E}_\xi\left(\min_{z\in Z(y)} c\,(y,\xi,z)\right). \tag{2}$$

Optimal solutions and the objective in (1) depend on the probability measure for ξ. It follows from this observation already that it may not make sense to solve problem (1) with ultimate precision, if the probability measure is not known precisely. A reasonable accuracy goal depends on the knowledge of the measure of ξ. Indeed, the underlying probability measure is not known precisely in many situations. This particularly holds in the following three cases:

(i) Often, the measure in (1) is not available explicitly. Instead, the empirical measure is employed, which is built from historically observed samples. The empirical measure is just an approximation of the true baseline measure, and one may not expect that the solution of problem (1) subject to the empirical measure is the optimal solution for the true baseline model, even if the sample size is large.

(ii) Many models of economic relevance artificially involve possible outcomes (often called scenarios) ξ to model potential future behavior of the economy. As the future behavior will differ from the artificially chosen scenarios (i.e., with probability one) it is necessary to have a framework which justifies the use of scenarios.

(iii) Numerical solution techniques typically replace the original probability measure in problem (1) by a discrete approximation. As for the empirical measure, the solution of (1) subject to the approximating measure is not more than an approximation.

This raises the general question if there are useful conditions to guarantee that the solutions, obtained by numerical schemes, are relevant for the initial problem (1) at all?

Different solution techniques for nonlinear stochastic optimization problems include sample average approximation methods (cf. Shapiro et al. [31]) and stochastic approximation methods (cf. Nemirovski et al. [18]), which employ samples, but not the probability measure directly. Solution techniques and heuristics from global optimization particularly apply to solve the general, nonlinear problem (1), and we refer to the extensive literature on global optimization.

This paper provides bounds for the objective of the genuine problem (1). The results are presented in terms of the Wasserstein distance. The distance provides a quality control of solutions of approximations of (1), as the objective turns out to be continuous with respect to the distance. Approximations, found by applying discrete measures, are justified in this way.

In addition we adapt Newton's method to the particular structure of the nonlinear stochastic optimization problem (1). Newton's procedure makes predictor-corrector methods available without additional effort, which we exploit to solve the inner problem for different scenarios faster. The results are finally demonstrated in an extensive case study. The case study addresses a stochastic extension of the optimal power flow problem taken from electrical engineering.

Outline of the paper. The following Section 2 addresses the theoretical justifications by discussing the baseline probability measure and its impact to problem (1). This section contains the central result of the paper, Theorem 5. Section 3 addresses numerical solution techniques, which are useful in reducing the computational burden to solve the nonlinear problem. The second part of the paper outlines the stochastic optimal power flow problem in Section 4. We conclude with results, a what-if analysis and a discussion in Section 5.

2 Approximation of the probability measure, and scenario reduction

This section introduces the Wasserstein distance for probability measures. It is demonstrated that the objective of the stochastic optimization problem (1) is continuous with respect to the Wasserstein distance and in addition, the risk functional $\mathcal{R}$ is Wasserstein continuous as well. This is the essential result in investigating the two-stage optimization problem, particularly for nonlinear problems.

The results provided ensure that passing to different, simpler probability distributions does not destroy a solution, the quality of the solution is kept to a reasonable extent. This section thus ensures that the solution of the problem with respect to a simpler, discrete probability measure is a good solution for the real problem as well. Important is just the quality of the approximation of the measure in the Wasserstein distance.

2.1 Wasserstein distance

The Wasserstein distance provides a distance of probability measures. We prove first that the expectation and risk functionals are continuous with respect to the Wasserstein distance. This is the essential property with various consequences for stochastic optimization:

(i) it is possible to derive bounds for (1) by comparing evaluations for different probability measures;

(ii) continuous probability measures can be replaced by discrete measures, which are eligible for numeric computations;

(iii) the computational burden for the numerical computation of the initial problem (1) is much lower for a smaller number of scenarios. The results justify clustering methods to reduce the computational burden, while error bounds are made available simultaneously.

Several books are dedicated to the Wasserstein distance. Details and mathematical properties of this distance can be found in the books by Rachev and Rüschendorf [28] or Villani [32].

Definition 1 (Wasserstein distance)**.** The Wasserstein distance of order $r \geq 1$ of two probability measures P and $\tilde{P}$ on the metric space (Ξ, d) is given by

$$\mathsf{d}_r(P, \tilde{P}) = \left(\inf \iint_{\Xi \times \Xi} d(\xi_1, \xi_2)^r \pi(\mathrm{d}\xi_1, \mathrm{d}\xi_2) \right)^{1/r},$$

where the infimum is among all probability measures π with marginals P and $\tilde{P}$, that is, they satisfy $\pi(A \times \Xi) = P(A)$ and $\pi(\Xi \times B) = \tilde{P}(B)$ whenever A and B are measurable sets.

Remark. The Wasserstein distance depends on the distance d on the initial space and is occasionally considered with a cost function instead of the distance d (cf. Villani [32]). For the applications in mind of this paper it is enough to consider $\Xi = \mathbb{R}^m$, equipped with a distance induced by a (weighted) Euclidean norm.

We have the following lemma for the pushforward measure (image measure) under Hölder, or Lipschitz continuous functions.

Lemma 2. *Let $Q : \left(\tilde{\Xi}, \tilde{d}\right) \to (\Xi, d)$ be a function between metric spaces which is Hölder continuous,*

$$d\big(Q(x), Q(y)\big) \leq C \cdot d(x, y)^\beta \tag{3}$$

with exponent $\beta \leq 1$. For β fixed, denote the infimum of the constants satisfying (3) by $\|Q\|_\beta$. Then

$$\mathsf{d}_r\left(P^Q, \tilde{P}^Q\right) \leq \|Q\|_\beta \cdot \mathsf{d}_{\beta r}(P, \tilde{P})^\beta,$$

where $P^Q(A) := P(Q \in A)$ is the image measure (pushforward measure) and $r \geq 1/\beta$.

Proof. Let π have marginals $\pi(A \times \Xi) = P(A)$ and $\pi(\Xi \times B) = \tilde{P}(B)$, then $\tilde{\pi}(A \times B) := \pi\left(Q^{-1}(A) \times Q^{-1}(B)\right)$ has marginals P^Q and $\tilde{P}^Q$. Hence, by the change of variables formula,

$$\begin{aligned} \mathsf{d}_r\left(P^Q, \tilde{P}^Q\right)^r &\leq \iint d(x, y)^r \tilde{\pi}(\mathrm{d}x, \mathrm{d}y) = \iint d\big(Q(x), Q(y)\big)^r \pi(\mathrm{d}x, \mathrm{d}y) \\ &\leq \iint \|Q\|_\beta^r \cdot d(x, y)^{\beta r} \pi(\mathrm{d}x, \mathrm{d}y). \end{aligned}$$

Taking the respective infimum reveals that

$$\mathsf{d}_r\left(P^Q, \tilde{P}^Q\right)^r \leq \|Q\|_\beta^r \cdot \mathsf{d}_{\beta r}\left(P, \tilde{P}\right)^{\beta r},$$

from which the assertion follows. □

Remark 3 (Curse of dimensionality)*.* There are some results characterizing the convergence of the empirical measure[1] $P_n := \frac{1}{n} \sum_{i=1}^n \delta_{\xi_i}$ towards its limit P in Wasserstein distance (cf. Graf and Luschgy [12] or Bolley et al. [3]). In rough words, convergence is of order $\mathcal{O}\left(n^{-1/m}\right)$, whenever $\Xi = \mathbb{R}^m$. Hence, the number of samples ξ has to be increased by a factor of 2^m, whenever an improvement of $\mathsf{d}(P_n, P)$ by a factor of $1/2$ is desired. In many situations of practical relevance the factor 2^m is much too high for numerical tractability.

Lemma 2 provides an essential reduction. By involving a Lipschitz, or Hölder continuous function function $Q : \mathbb{R}^{m_1} \to \mathbb{R}^{m_2}$ with $m_2 \ll m_1$ the measures P_n^Q converge much faster towards P^Q, the order is $\mathcal{O}\left(n^{-1/m_2}\right)$.

[1] δ_ξ is the Dirac measure (or point measure) at ξ, i.e., $\delta_\xi(A) = \begin{cases} 1 & \text{if } \xi \in A \\ 0 & \text{else.} \end{cases}$

2.2 Coherent risk functionals

The risk functionals considered in problem (1) for $\mathbb{R}$-valued random variables Q are of the form

$$\mathcal{R}(Q) = \sup_{\sigma \in \mathcal{S}} \mathcal{R}_\sigma(Q), \tag{4}$$

where $\mathcal{R}_\sigma(Q) := \int_0^1 F_Q^{-1}(u)\sigma(u)\mathrm{d}u$ is called a *distortion risk functional* (or *spectral risk functional*) and $\sigma \in \mathcal{S}$ is a distortion function: a distortion function $\sigma : [0,1) \to [0,\infty)$ is nonnegative, nondecreasing and satisfies $\int_0^1 \sigma(u)\mathrm{d}u = 1$. $F_Q^{-1}(\alpha) := \inf\{q\colon P(Q \le q) \ge \alpha\}$ is the quantile.

The expectation is a simple version of a risk functional (4), as

$$\mathbb{E}\, Q = \int_0^1 F_Q^{-1}(u)\mathrm{d}u.$$

A specific example, which is often employed in stochastic optimization, is the (upper) Average Value-at-Risk (or Conditional Value-at-Risk) at level α defined as

$$\mathsf{AV@R}_\alpha(Q) := \frac{1}{1-\alpha}\int_\alpha^1 F_Q^{-1}(u)\mathrm{d}u.$$

These types of coherent risk functionals are discussed in many places, not only in stochastic optimization, but particularly in mathematical finance. For a broad discussion on risk functionals we refer to the book [24] by Pflug and Römisch, and for distortion risk functionals in particular to Pflug [20]. A comprehensive mathematical treatment can be found in Pichler [26, 25] as well.

It is of essential importance for stochastic optimization that the risk functionals (4) are continuous with respect to the Wasserstein distance. This is the content of the following lemma.

Lemma 4. *Distortion risk functionals are continuous with respect to changing the underlying probability measure, that is,*

$$\left|\mathcal{R}_{\sigma;P}(Q) - \mathcal{R}_{\sigma;\tilde{P}}(Q)\right| \le \|Q\|_\beta \cdot \mathsf{d}_{\beta p}(P,\tilde{P})^\beta \cdot \|\sigma\|_q\,.$$

$q \ge 1$ is the exponent conjugate to p, $\frac{1}{p}+\frac{1}{q}=1$.

Proof. It follows from Hölder's inequality that

$$\begin{aligned}
\mathcal{R}_{\sigma;P}(Q) - \mathcal{R}_{\sigma;\tilde{P}}(Q) &= \int_0^1 \left(F_{Q;P}^{-1}(u) - F_{Q;\tilde{P}}^{-1}(u)\right)\sigma(u)\mathrm{d}u \\
&\le \left(\int_0^1 \left|F_{Q;P}^{-1}(u) - F_{Q;\tilde{P}}^{-1}(u)\right|^p \mathrm{d}u\right)^{1/p} \left(\int_0^1 \sigma(u)^q \mathrm{d}u\right)^{1/q} \\
&= \left(\int_0^1 \left|F_{Q;P}^{-1}(u) - F_{Q;\tilde{P}}^{-1}(u)\right|^p \mathrm{d}u\right)^{1/p} \|\sigma\|_q\,,
\end{aligned}$$

where the probability measure P is explicitly exposed as a subscript by writing $F_{Q;P}^{-1}(\alpha) = \inf\{q \in \mathbb{R} :\ P(Q \le q) \ge \alpha\}$. Due to the identity

$$\int_0^1 \left|F_{Q;P}^{-1}(u) - F_{Q;\tilde{P}}^{-1}(u)\right|^p \mathrm{d}u = \mathsf{d}_p\left(P^Q, \tilde{P}^Q\right)^p$$

in Ambrosio et al. [1, Theorem 6.0.2] it follows from Lemma 2 that

$$\mathcal{R}_{\sigma;P}(Q)-\mathcal{R}_{\sigma;\tilde{P}}(Q)\le\|Q\|_\beta\cdot\|\sigma\|_q\cdot\mathsf{d}_{\beta p}(P,\tilde{P})^\beta.$$

The assertion is immediate by interchanging the roles of P and $\tilde{P}$. □

The following theorem combines the ingredients collected, it is the central statement of the present text. The theorem states that the stochastic optimization problem (1) is continuous with respect to the Wasserstein distance.

Theorem 5 (Continuity of the stochastic optimization problem (1))**.** *Let c be uniformly Hölder continuous in its random component ξ, that is*

$$c(y,\xi,z)-c(y,\tilde{\xi},z)\le\|c\|_\beta\cdot d\big(\xi,\tilde{\xi}\big)^\beta$$

for all y, z, ξ and $\tilde{\xi}$. Then the stochastic optimization problem (1) is continuous in its probability measure, it satisfies

$$\left|\inf_{y\in Y}\mathcal{R}_P\left(\inf_{z\in Z}c(y,\xi,z)\right)-\inf_{y\in Y}\mathcal{R}_{\tilde{P}}\left(\inf_{z\in Z}c(y,\xi,z)\right)\right|\le\|c\|_\beta\cdot\mathsf{d}_{\beta p}(P,\tilde{P})^\beta\cdot\sup_{\sigma\in\mathcal{S}}\|\sigma\|_q\,. \tag{5}$$

Remark 6. The inequality (5) provides an upper bound on the error induced when replacing a probability measure P by $\tilde{P}$. We employ this result later and approximate P by a simple probability measure $\tilde{P}$. The theorem provides a bound then in both directions when comparing the solution of the approximating problem $(\tilde{P})$ with the true problem (P).

Proof of Theorem 5. The infimum of uniformly Hölder continuous functions is Hölder continuous again (this is detailed in Lemma 21 in the Appendix), hence

$$Q(y,\xi):=\inf_{z\in Z}c(y,\xi,z)$$

is Hölder continuous with the same constant $\|c\|_\beta$. It follows from Lemma 4 that

$$\mathcal{R}_{\sigma;P}(Q)-\mathcal{R}_{\sigma;\tilde{P}}(Q)\le\|c\|_\beta\cdot\|\sigma\|_q\cdot\mathsf{d}_{\beta p}(P,\tilde{P})^\beta$$

for every $\sigma\in\mathcal{S}$.

Now choose $\sigma_\varepsilon\in\mathcal{S}$ such that $\sup_{\sigma\in\mathcal{S}}\mathcal{R}_{\sigma;P}(Q)\le\mathcal{R}_{\sigma_\varepsilon;P}(Q)+\varepsilon$, and it becomes obvious that

$$\mathcal{R}_P(Q)-\mathcal{R}_{\tilde{P}}(Q)-\varepsilon\le\mathcal{R}_{\sigma_\varepsilon;P}(Q)-\mathcal{R}_{\sigma_\varepsilon;\tilde{P}}(Q)\le\|c\|_\beta\cdot\sup_{\sigma\in\mathcal{S}}\|\sigma\|_q\cdot\mathsf{d}_{\beta p}(P,\tilde{P})^\beta.$$

By the same reasoning as above it follows that

$$\inf_{y\in Y}\mathcal{R}_P\big(Q(y,\xi)\big)-\inf_{y\in Y}\mathcal{R}_{\tilde{P}}\big(Q(y,\tilde{\xi})\big)\le\|c\|_\beta\cdot\sup_{\sigma\in\mathcal{S}}\|\sigma\|_q\cdot\mathsf{d}_{\beta p}(P,\tilde{P})^\beta.$$

This is the assertion, as the roles of P and $\tilde{P}$ can be interchanged. □

Remark 7 (Integer and binary variables in the constraints)*.* Theorem 5 provides sufficient conditions for continuity of the general stochastic optimization problem (1) with respect to changing the probability measure. It is worth a remark that the infimum in Theorem 5 is among general sets Y in the first stage and general sets Z in the second stage, no special structure of these sets is required. This means in particular that the outer and inner minimization can be over integers or may include binary variables, and the conclusion on continuity is still valid.

2.3 Scenario reduction and clustering

Every probability measure P on $\Xi=\mathbb{R}^m$ can be approximated arbitrarily close in the Wasserstein distance by a discrete measure $\tilde{P}=\sum_{i=1}^{\tilde{n}}\tilde{p}_i\,\delta_{\xi_i}$ with $\tilde{p}_i>0$ and $\xi_i\in\Xi$. It follows from Theorem 5 that every stochastic optimization problem (1) can be approximated by replacing the (eventually continuous) measure P by a simple, discrete measure $\tilde{P}$. In this way, every stochastic optimization problem is eligible for computational, numerical treatment.

Using clustering one intends to replace a probability measure by a simpler one, such that the computation of (1) can be done even more quickly and more efficiently. In what follows we demonstrate that clustering is continuous with respect to the Wasserstein distance, which makes clustering a useful method in reducing the computational burden.

To this end consider a discrete probability measure

$$\tilde{P}=\sum_{i=1}^{\tilde{n}}\tilde{p}_i\,\delta_{\tilde{\xi}_i},$$

where $\tilde{\Xi}:=\left\{\tilde{\xi}_i\colon i=1,\dots\tilde{n}\right\}$ are the supporting points. One may compare this measure with P.

The following Definition and Lemma (cf. Pflug and Pichler [21]) provides the tool to determine the discrete measure $\tilde{P}$, which approximates the original measure P as well as possible in terms of the Wasserstein distance.

Definition 8. A *tessellation* of Ξ consists of measurable sets $(V_i)_{i=1}^{\tilde{n}}$ such that

$$\bigcup_{i=1}^{\tilde{n}}V_i=\Xi\quad\text{ and }\quad V_i\cap V_j=\emptyset\text{ whenever }i\neq j.$$

A tessellation is a *Voronoi tessellation* with centers $\left\{\tilde{\xi}_i:i=1,\dots\tilde{n}\right\}$, if

$$d(\xi,\tilde{\xi}_i)\leq d(\xi,\tilde{\xi}_k)\text{ for all }\xi\in V_i\text{ and }k=1,\dots\tilde{n}.$$

Lemma 9. *The probability measure $\tilde{P}=\sum_{i=1}^{\tilde{n}}\tilde{p}_i\,\delta_{\tilde{\xi}_i}$, which is located on $\left\{\tilde{\xi}_i:i=1,\dots\tilde{n}\right\}$ and approximating P in best possible way in terms of the Wasserstein distance has the weights*

$$\tilde{p}_i:=P(V_i),\tag{6}$$

where $(V_i)_{i=1}^{\tilde{n}}$ is a Voronoi tessellation with centers $\left\{\tilde{\xi}_i:i=1,\dots\tilde{n}\right\}$. The best distance is given by the explicit formula

$$\mathsf{d}_r\left(P,\sum_{i=1}^{\tilde{n}}\tilde{p}_i\delta_{\tilde{\xi}_i}\right)^r=\int_\Xi\min_{j=1,\dots\tilde{n}}d(\xi,\tilde{\xi}_j)^r\,P(\mathrm{d}\xi)\tag{7}$$

for all $r\geq1$.

Remark. It is a consequence of Lemma 9 and (7) that the Voronoi tessellation does not have to be available explicitly in order to compute the Wasserstein distance if the weights are chosen as specified in (6). For algorithms to compute (7) or approximations of it we may refer to Pflug and Pichler [22, 23]. These reference address the problem of finding optimal locations $\tilde{\xi}_1,\dots\ \tilde{\xi}_{\tilde{n}}$ as well by employing clustering methods and stochastic approximation.

Proof of Lemma 9. Let π have marginals P and $\tilde{P}$, then

$$\iint d(\xi,\tilde{\xi})^r \pi(\mathrm{d}\xi,\mathrm{d}\tilde{\xi}) \geq \iint \min_{j=1,\dots\tilde{n}} d(\xi,\tilde{\xi}_j)^r \pi(\mathrm{d}\xi,\mathrm{d}\tilde{\xi}) = \int \min_{j=1,\dots\tilde{n}} d(\xi,\tilde{\xi}_j)^r P(\mathrm{d}\xi),$$

and hence $\mathsf{d}_r(P,\tilde{P})^r \geq \int \min_{j=1,\dots\tilde{n}} d(\xi,\tilde{\xi}_j)^r\, P(\mathrm{d}\xi)$, a lower bound.

Next, define the map $T(\xi) := \tilde{\xi}_i,\ $ if $\xi \in V_i$ and the bivariate measure $\pi(A\times B) := P\left(A \cap T^{-1}(B)\right)$. It holds that $\pi(A\times\Xi) = P(A)$ and $\pi(\Xi \times B) = P^T(B) = \tilde{P}(B)$, such that π has adjusted marginals. It thus holds that

$$\iint d(\xi,\tilde{\xi})^r \pi(\mathrm{d}\xi,\mathrm{d}\tilde{\xi}) = \iint \min_{j=1,\dots\tilde{n}} d(\xi,\tilde{\xi}_j)^r \pi(\mathrm{d}\xi,\mathrm{d}\tilde{\xi}) = \int \min_{j=1,\dots\tilde{n}} d(\xi,\tilde{\xi}_j)^r P(\mathrm{d}\xi),$$

such that

$$\mathsf{d}_r\left(P,\sum_{i=1}^{\tilde{n}} \tilde{p}_i \delta_{\tilde{\xi}_i}\right)^r = \int \min_{j=1,\dots\tilde{n}} d(\xi,\tilde{\xi}_j)^r P(\mathrm{d}\xi),$$

the assertion. □

3 Implication on Algorithms to solve nonlinear stochastic problems numerically

Theorem 5 in the previous section gives a constructive bound for the objective of the two-stage stochastic optimization problem, whenever the probability measure P in (1) is replaced by a simpler measure $\tilde{P}$. Lemma 9, in addition, gives an explicit formula for the Wasserstein distance of the probability measure, which is located on $\left\{\tilde{\xi}_i:\ i=1,\dots n\right\}$.

These results have the following implications for numerical solutions:

(i) A continuous probability measure is not eligible for numerical computations in (1). However, numerical algorithms easily apply for the problem with probability measure P replaced by $\tilde{P} = \sum_{i=1}^{\tilde{n}} p_i\, \delta_{\tilde{\xi}_i}$. Hence, even the continuous problem gets numerically tractable, and an explicit bound is given by (5) in Theorem 5 (cf. also Remark 6).

(ii) A further consequence is that complicated problems involving a measure of the form $\sum_{i=1}^{n} p_i\, \delta_{\xi_i}$ can be replaced by a simpler measure $\tilde{P} = \sum_{i=1}^{\tilde{n}} p_i\, \delta_{\tilde{\xi}_i}$, where $\tilde{n}$ is much smaller than n, $\tilde{n} \ll n$. This reduces numerical computation times significantly (cf. Lemma 9).

For each realization of $\tilde{\xi}_i$, $i = 1,\dots\tilde{n}$, the inner minimization of the stochastic optimization problem (1) has to be solved separately. The following section addresses and exploits this particular problem structure.

3.1 Separability and decomposition

We consider a discrete realization for the measure $P = \sum_{i=1}^{n} p_i\, \xi_i$ (for example an empirical measure resulting from Monte Carlo simulations, or a measure reduced according Lemma 9). For every realization $\xi \in \Xi := \{\xi_i\colon\ i = 1,\dots n\}$ the recourse problem

$$\min_{z\in Z(y)} c(y,\xi,z),$$

i.e., the inner problem in (1), has to be solved. Without loss of generality one may assume that the problem is given in the form

$$\begin{aligned} Q(\xi) := \quad & \text{minimize}_{\text{ in } z}\, c(\xi, z) \\ & \text{subject to } h(\xi, z) = 0, \\ & z \in \mathrm{B}. \end{aligned} \tag{8}$$

Q is called the *recourse* function. The essential observation is that every scenario ξ can be considered as a parameter, and $Q(\xi)$ then are the final, total costs after optimization for the particular scenario ξ. The optimal solution z notably differs for every fixing of ξ, such that the solution z of (8) is a function of ξ, $z = z(\xi)$.

Feasibility. To evaluate the expected value or $\mathcal{R}$ (cf. Eq. (1)), the recourse function $Q(\xi)$ has to be evaluated for every assignment of ξ, that is, the inner minimization in z (Eq. (8)) has to be feasible for every outcome of ξ separately. Indeed, if Q were not feasible almost everywhere, then $P(Q = \infty) > 0$ and thus $\mathcal{R}(Q) \geq \mathbb{E}\, Q = \infty$.

We list this following, important observation regarding the feasibility of the problem for different scenarios.

Lemma 10. *Let Ξ and B be compact and the constraint function h be continuous. Then the set of feasible scenarios $\{\xi \in \Xi :\ \exists z \in \mathrm{B} : h(\xi, z) = 0\}$ is closed and compact.*

Proof. Consider the set $F := \{(\xi, z) \in \Xi \times \mathrm{B} :\ h(\xi, z) = 0\}$. This set is closed, as h is continuous and $F = h^{-1}(\{0\})$. The set F is moreover compact, as $F \subset \Xi \times \mathrm{B}$ and Ξ and B are compact. The projection $i : \Xi \times \mathrm{B} \to \Xi$ is continuous, such that $\{\xi \in \Xi :\ \exists z \in \mathrm{B} : h(\xi, z) = 0\} = i(F)$ is compact. □

Remark 11. Especially in an economic situation it is typically expected that the recourse function $Q(\xi)$ provides similar results for parameters ξ, which are close (i.e., that Q is continuous). However, the previous Lemma 10 explains that this cannot be expected in general. It is an important consequence of Lemma 10 that small aberrations from a feasible scenario ξ may *not* be feasible for the inner problem (8). As a consequence the result of the problem (8) is possibly not continuous with respect to the parameter ξ, such that further conditions are necessary to apply Theorem 5. Another consequence is that passing to a simpler measure, as outlined in Lemma 9, may not be possible.

Reasonable conditions to insure the assertions of Theorem 5 and Lemma 9 often can be derived from the particular problem at hand, sometimes in connection with the Remarks 13 or 14 below.

3.2 Numerical solutions by employing Newton–Raphson

The nonlinear optimization problem (8) is typically (and efficiently) solved by applying Newton's method.[2] For this the Lagrangian $L(z, \lambda; \xi) := c(\xi, z) + \lambda^\top h(\xi, z)$ is considered, where we treat ξ as a parameter.

[2] To simplify the exposition and for convenience of notation we treat active inequality constraints and the box constraints $z \in \mathrm{B}$ as equations incorporated in $h(\xi, z) = 0$ (cf. Ruszczyński [29, p. 326]). The active constraints notably vary with ξ and z.

The necessary conditions of optimality are

$$f(z,\lambda;\xi) := \begin{pmatrix} L_z(z,\lambda;\xi) \\ L_\lambda(z,\lambda;\xi) \end{pmatrix} = \begin{pmatrix} c_z(\xi,z) + \lambda^\top h_z(\xi,z) \\ h(\xi,z) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{9}$$

where $c_z(\xi,z) = \nabla_z c(\xi,z) = \left(\frac{\partial c}{\partial z_1}, \dots \frac{\partial c}{\partial z_n}\right)$ is the partial derivative of c with respect to the vector z (h_z is the partial derivative of h, resp.), etc.

The symmetric Jacobian of the system of equations (9),

$$f'(z,\lambda;\xi) = \begin{pmatrix} c_{zz}(\xi,z) + \lambda^\top h_{zz}(\xi,z) & h_z(\xi,z)^\top \\ h_z(\xi,z) & 0 \end{pmatrix}, \tag{10}$$

is often called KKT-system ($c_{zz} = \nabla_z^2 c$ is the symmetric Jacobian matrix with entries $\frac{\partial^2 c}{\partial z_i \partial z_j}$, etc.).

Starting with some tentative solution $\begin{pmatrix} z_0 \\ \lambda_0 \end{pmatrix}$ to solve (9), Newton's method provides the iterates

$$\begin{pmatrix} z_{k+1} \\ \lambda_{k+1} \end{pmatrix} := \begin{pmatrix} z_k \\ \lambda_k \end{pmatrix} + \begin{pmatrix} \Delta z_k \\ \Delta \lambda_k \end{pmatrix},$$

where the linear system of equations

$$f'(z_k,\lambda_k;\xi) \cdot \begin{pmatrix} \Delta z_k \\ \Delta \lambda_k \end{pmatrix} = -f\left(z_k,\lambda_k;\xi\right)$$

has to be solved in successive iterations.

Remark 12. The Lagrange multiplier λ is uniquely determined whenever the matrix $h_z(\xi,z)$ has linearly independent rows. Further, regularity of the matrix (10) can often be verified by employing Cauchy's interlacing eigenvalue theorem. This theorem ensures that the eigenvalues of (10) are strictly positive or strictly negative, such that the matrix is regular (invertible).

Details on Newton's procedure are elaborated, e.g., in Ruszczyński [29] and in Boyd and Vandenberghe [4].

Remark 13. The implicit function theorem provides sufficient conditions to ensure that h is invertible in a neighborhood. When employing Newton's method to solve problem (8), then relevant information is automatically available due to the (inverted) Jacobian matrix. This provides numerical evidence for a feasible region in a neighborhood of a single, feasible scenario.

3.3 Predictor

Newton's methods converges quickly whenever a good starting value is available. In order to obtain a good starting value a predictor corrector method can be employed.[3] The predictor provides a reasonable, tentative guess, and the corrector improves the initial guess to compute a solution. It turns out that a predictor, as well as a corrector, are provided by Newton's method.

To compute $z(\tilde{\xi})$ one may consider $z(\xi + \Delta\xi)$, where $\Delta\xi := \tilde{\xi} - \xi$. Provided that $\Delta\xi$ is small it is to be expected that $z(\xi) + z'(\xi) \cdot \Delta\xi$ is a reasonable starting value for Newton's method, whenever

[3] The idea of predictor corrector methods is adopted from evolution equations (differential equations).

$z(\xi)$ is known. This predictor can be specified by taking the derivative of (9) with respect to the parameter ξ, resulting in the equations

$$\begin{pmatrix} L_{zz} & L_{\lambda z}^\top \\ L_{\lambda z} & 0 \end{pmatrix} \cdot \begin{pmatrix} z_\xi \\ \lambda_\xi \end{pmatrix} + \begin{pmatrix} L_{z\xi} \\ L_{\lambda\xi} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

This system rewrites as

$$\begin{pmatrix} c_{zz}(\xi,z) + \lambda^\top h_{zz}(\xi,z) & h_z(\xi,z)^\top \\ h_z(\xi,z) & 0 \end{pmatrix} \cdot \begin{pmatrix} z_\xi \\ \lambda_\xi \end{pmatrix} + \begin{pmatrix} c_{z\xi}(\xi,z) + \lambda^\top h_{z\xi}(\xi,z) \\ h_\xi(\xi,z) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

For an incremental change of $\Delta\xi$ we have that $\Delta z = z' \cdot \Delta\xi$, and the latter equation thus is

$$\begin{pmatrix} c_{zz}(\xi,z) + \lambda^\top h_{zz}(\xi,z) & h_z(\xi,z)^\top \\ h_z(\xi,z) & 0 \end{pmatrix} \cdot \begin{pmatrix} \Delta z \\ \Delta\lambda \end{pmatrix} = -\begin{pmatrix} c_z(\xi+\Delta\xi,z) + \lambda^\top h_z(\xi+\Delta\xi,z) \\ h(\xi+\Delta\xi,z) \end{pmatrix}, \tag{11}$$

because (9) holds with equality at ξ.

The linear equation (11) has to be solved to obtain the predictor $z(\xi)+\Delta z$. But this equation is notably the equation for the Newton step at $z(\xi)$, except that the right hand side is disturbed. Hence, starting Newton's method at ξ by employing the solution $z(\xi)$ represents a predictor in direction of the Taylor approximation to compute $z(\tilde{\xi})$. Importantly, an implementation of Newton's method can be *reused* to numerically compute the predictor without additional effort.

This feature exposes Newton's method as a central tool to compute the recourse function Q of the stochastic optimization problem (1) for varying scenarios.

3.4 The envelope theorem

Exact knowledge of $z(\xi)$ is not necessary to compute the recourse function $Q(\xi)$. Even more, in many situations a reliable bound for $Q(\xi)$ might be enough for some selected scenarios ξ or for scenarios with small probability p_ξ. Given that we can estimate an upper bound L for the derivative it holds that $Q(\tilde{\xi}) \geq Q(\xi) - L\left\|\xi - \tilde{\xi}\right\|$ and $Q(\tilde{\xi}) \leq Q(\xi) + L\left\|\xi - \tilde{\xi}\right\|$. The envelope theorem, which we address in what follows, often provides such a reasonable bound for the derivative.

Consider the optimization problem[4]

$$\begin{aligned} Q(\xi) := \quad & \text{minimize }_{\text{in } z}\ c(\xi,z) \\ & \text{subject to }\ h(\xi,z) = 0, \end{aligned} \tag{12}$$

where we treat ξ as a parameter. An application of the envelope theorem provides the derivative $Q_\xi(\xi)$ as a simple byproduct of the optimization (12). Combined with the Taylor series expansion we obtain the approximation $Q(\tilde{\xi}) \simeq Q(\xi) + Q_\xi(\xi)\cdot(\tilde{\xi}-\xi)$. An approximation of $Q(\tilde{\xi})$ thus is available, without explicitly solving (12) for the optimal value $z(\tilde{\xi})$. This approximation is sufficient for many scenarios $\tilde{\xi}$, especially if $\tilde{\xi}$ is close to ξ, or if its corresponding probability $p_{\tilde{\xi}}$ is small.

The optimal value z in (12) is a function of ξ, which satisfies the implicit conditions

$$\begin{aligned} Q(\xi) &= c\big(\xi, z(\xi)\big) \qquad \text{and} \\ h\big(\xi, z(\xi)\big) &= 0. \end{aligned}$$

[4] Again, active box constraints are incorporated in h, cf. Footnote 2 on page 9.

By differentiating these equations it follows that

$$\begin{aligned} Q_\xi(\xi) &= c_\xi\big(\xi, z(\xi)\big) + c_z\big(\xi, z(\xi)\big)\cdot z'(\xi) \quad \text{and}\\ 0 &= h_\xi\big(\xi, z(\xi)\big) + h_z\big(\xi, z(\xi)\big)\cdot z'(\xi). \end{aligned} \tag{13}$$

Moreover, employing the Lagrangian $L(z,\lambda) = c(\xi, z) + \lambda^\top h(\xi, z)$ for (12) the first order conditions are

$$\begin{aligned} L_z(z,\lambda) &= c_z(\xi, z) + \lambda^\top h_z(\xi, z) = 0 \quad \text{and} \\ L_\lambda(z,\lambda) &= h(\xi, z) = 0. \end{aligned} \tag{14}$$

After multiplying (13) with $\lambda^\top$ and (14) with $z'(\xi)$ it becomes evident that $\lambda^\top h_\xi\big(\xi, z(\xi)\big) = c_z\big(\xi, z(\xi)\big)\cdot z'(\xi)$. Hence,

$$\begin{aligned} Q_\xi(\xi) &= c_\xi\big(\xi, z(\xi)\big) + c_z\big(\xi, z(\xi)\big)\cdot z'(\xi)\\ &= c_\xi\big(\xi, z(\xi)\big) + \lambda^\top h_\xi(\xi, z), \end{aligned} \tag{15}$$

in accordance which the envelope theorem.

Remark 14 (Lipschitz constant of the recourse function). It follows from (15) (the assertion of the envelope theorem) that

$$\|Q_\xi\| \le \|c_\xi\| + \left\|\lambda^\top h_\xi\right\|.$$

The Lipschitz constant of the recourse function Q thus is bounded by

$$\operatorname{Lip}(Q) \le \|c_\xi\| + \left\|h_\xi^\top\right\|\cdot\|\lambda\|,$$

where $\left\|h_\xi^\top\right\|$ is the consistent matrix norm induced by the norm for $\|\lambda\|$. The norm of c_ξ is available by inspecting the cost function.

Boundedness of $\left\|h_\xi^\top\right\|$ often can be derived for the particular problem at hand, while boundedness of the dual variable $\|\lambda\|$ is addressed in Remark 12 above. In addition the dual variable $\lambda(\xi)$ can be monitored during the computation of $Q(\xi)$ to track its norm, such that a global Lipschitz constant of the recourse function Q is available during the computations.

3.5 2$^{\text{nd}}$ stage speed-up

It was elaborated that Newton's method can be employed to evaluate the recourse function $Q(\xi)$ for different parameters ξ by minimizing with respect to z. Further, Newton's method automatically provides a predictor by starting at a previous solution.

We recall here two variants of Newton's method to exploit this method further and to accelerate numerical computations.

Quasi Newton, or Newton-like methods. Newton's method requires computing the Newton step, that is, the matrix in (10) has to be inverted. As it is time consuming to invert these matrices in each iterative step it is tempting to reuse the (inverse) Jacobian matrices, or its LU or QR decomposition from previous iterations. Indeed, the Jacobian can be reused as long as $\|f(z_k, \xi)\|$ decreases during an iteration.

The (inverse) Jacobian thus can be reused in both situations,

(i) during the computation of $Q(\xi)$ when solving the system $f(z,\xi)=0$ with respect to z, and

(ii) as predictor to restart the Newton procedure to minimize $Q(\tilde{\xi})$ for a new scenario $\tilde{\xi}$, which is different (but in ideal case close) to the previous ξ.

It is the advantage of reusing the inverse Jacobian that it does not have to be constructed nor inverted again, which is typically the most expensive step. However, convergence is not further insured, or can be expected to be slower. Typically linear convergence is obtained, whereas quadratic convergence of the full Newton method is lost.

A survey on the convergence of Newton-like methods can be found in Yamamoto [35]. The following theorem, dating back to Dennis [8], justifies the method outlined.

Theorem 15 (Convergence of Newton-like methods)**.** *Let $f \in C^2$ be twice continuously differentiable with $\|f'(z)-f'(\tilde{z})\| \le c \cdot \|z-\tilde{z}\|$ and let $M(z)$ satisfy $\|1-f'(z)M(z)\| \le \delta < 1$ and $\|M(z)\| \le B$. If z_0 can be chosen such that $\|f(z_0)\| < 2\frac{1-\delta}{B^2 c}$, then the Newton-like sequence*

$$z_{k+1} := z_k - M(z_k) \cdot f(z_k)$$

converges to a zero of f. Convergence is at least linearly.

Remark 16. The convergence result on Newton-like methods in Theorem 15 provides a qualitative alternative to the implicit function theorem (cf. Remark 11): given that the conditions in Theorem 15 are satisfied for the function $f(\cdot,\xi)$ for some ξ and z is available, such that $f(z,\xi)=0$, then, by assuming enough smoothness, $f(\cdot,\tilde{\xi})$ is invertible as well and Newton's method converges to $\tilde{z}$, the solution of $f(\tilde{z},\tilde{\xi})$.

Broyden's method. Although convergence is already obtained by employing the matrix M_k, which is available from a previous iteration, the speed, and the rate of convergence can be improved by successively updating the matrix M_k (cf. Broyden and Vespucci [5]). The following proposition provides a useful method to update the matrix M_k after each iteration.

Proposition 17. *Let M_k be a matrix and f_k^* an arbitrary functional such that $f_k^*(f_{k+1}-f_k) \neq 0$. Then the updated matrix*

$$M_{k+1} := M_k + \Big(z_{k+1} - z_k - M_k(f_{k+1}-f_k)\Big) \otimes \frac{f_k^*}{f_k^*(f_{k+1}-f_k)}$$

satisfies

$$M_{k+1}(f_{k+1}-f_k) = z_{k+1} - z_k \tag{16}$$

($\otimes$ is the outer product of two vectors).

Remark 18. Notice, that $f(z_{k+1}) \approx f(z_k) + f'(z_k)\cdot(z_{k+1}-z_k)$ by Taylor's expansion, and $z_{k+1}-z_k \approx f'(z_k)^{-1}(f_{k+1}-f_k)$ in first Tayler approximation. M_k thus can be expected to be an approximation of the inverse Jacobian, that is, $f'(z_k)^{-1} \approx M_k$. The new approximation M_{k+1} is an improved approximation of $f'(z_k)^{-1}$, at least it correctly recovers the new direction $z_{k+1}-z_k$.

Proof. The assertion of Proposition 17 is immediate, as

$$\begin{aligned} M_{k+1}(f_{k+1}-f_k) &= M_k(f_{k+1}-f_k) + \Big(z_{k+1}-z_k - M_k(f_{k+1}-f_k)\Big) \otimes \frac{f_k^*(f_{k+1}-f_k)}{f_k^*(f_{k+1}-f_k)} \\ &= M_k(f_{k+1}-f_k) + \big(z_{k+1}-z_k - M_k(f_{k+1}-f_k)\big) \\ &= z_{k+1}-z_k. \end{aligned}$$

Algorithm 1 Newton-like method to compute the recourse Q for all scenarios ξ

(i) **Initialization.** Choose an arbitrary $\xi_1 \in \Xi$, and set $M_1 := f'(\xi_1, z_1)^{-1}$. Select $0 < \beta < 1$.

(ii) **Newton Iteration**, k**.** Accept

$$z_{k+1} := z_k - M_k \cdot f(\xi_i, z_k) \tag{17}$$

and set

$$M_{k+1} := M_k + \Big(y_{k+1} - y_k - M_k(f_{k+1} - f_k)\Big) \otimes \frac{(f_{k+1} - f_k)^*}{\|f_{k+1} - f_k\|_2^2}$$

if

$$\|f(\xi_i, z_{k+1})\| < \beta \cdot \|f(\xi_i, z_k)\| \,. \tag{18}$$

If not, then employ directional search or restart with $M_k := f'(z_i, \xi_k)^{-1}$ until (18) holds again. Set $k \leftarrow k + 1$ and repeat the Newton iteration (17) unless

$$\|f(\xi_i, z_{k+1})\| < \varepsilon,$$

the desired precision goal, and continue with the next scenario (**Scenario Iteration** (iii)).

(iii) **Scenario Iteration,** i**.** Set $i \leftarrow i + 1$ and choose a scenario ξ_{i+1} which is *as close as possible* to ξ_i and for which $Q(\xi_{i+1})$ has not been computed yet. Continue the **Newton Iteration** (ii) by reusing the current matrix M_k.

□

The updated iterate M_{k+1} notably satisfies equation (16), which is the equation for the Newton–Raphson step. It is hence to be expected that the matrix M_{k+1} is a better approximation of $F(z_k)^{-1}$ than M_k. This is indeed the case, if f_k^* is chosen appropriately. For the choice $f_k^*(\cdot) := \langle f_{k+1} - f_k, \cdot\rangle$ convergence is even superlinear (cf. Gay [11]).

Remark 19. While f' and f'^{-1} are symmetric matrices (cf. (10)), the update proposed in Remark 18 is not necessarily symmetric again. Moreover the sparse structure of the matrix f' may be lost by the updates. The matrix can be kept symmetric by choosing the linear functional

$$f_k^*(\cdot) := \langle z_{k+1} - z_k - M_k(f_{k+1} - f_k), \cdot\rangle \,.$$

We refer to Nocedal [19] for further aspects on this topic.

Algorithm 1 outlines the techniques collected to accelerate the computation of all second stages.

4 Case study: electrical load flow

Bienstock [2] points out that power flow problems can surprise optimization experts by their difficulty. For this reason they are often considered as benchmark problems, in particular for the choice of algorithms to numerically solve optimization problems. From practical perspective, power flow problems are motivated by an increasing demand of electrical power and varying operational costs, as well as the need to manage system failures as blackouts (cf. Bienstock [2]).

A second motivation – with increasing importance – is the need to incorporate renewable energy in the power flow network: many countries are in a transition period and currently redesign their power network. Germany, for example, builds new transmission lines in order to transport electricity from the north, where electricity is generated in off-shore wind parks, to the south, where demand is high. A new situation is created in Denmark as well (cf. Villumsen et al. [34]), as the country has decided that nearly 50 % of its demand should be provided by wind power within a period of ten years. These situations provide the opportunity to question and rethink the existing power transmission network, and to improve its efficiency by establishing an optimal power network topology.

This case study addresses the optimal power flow problem in an economic, stochastic environment. The stochastic character is given by the fact that future demand and supply are random, they can only be assumed or estimated from today's perspective. However, the power grid has to be designed today, although its future utilization and the capacities necessary are not completely known today.

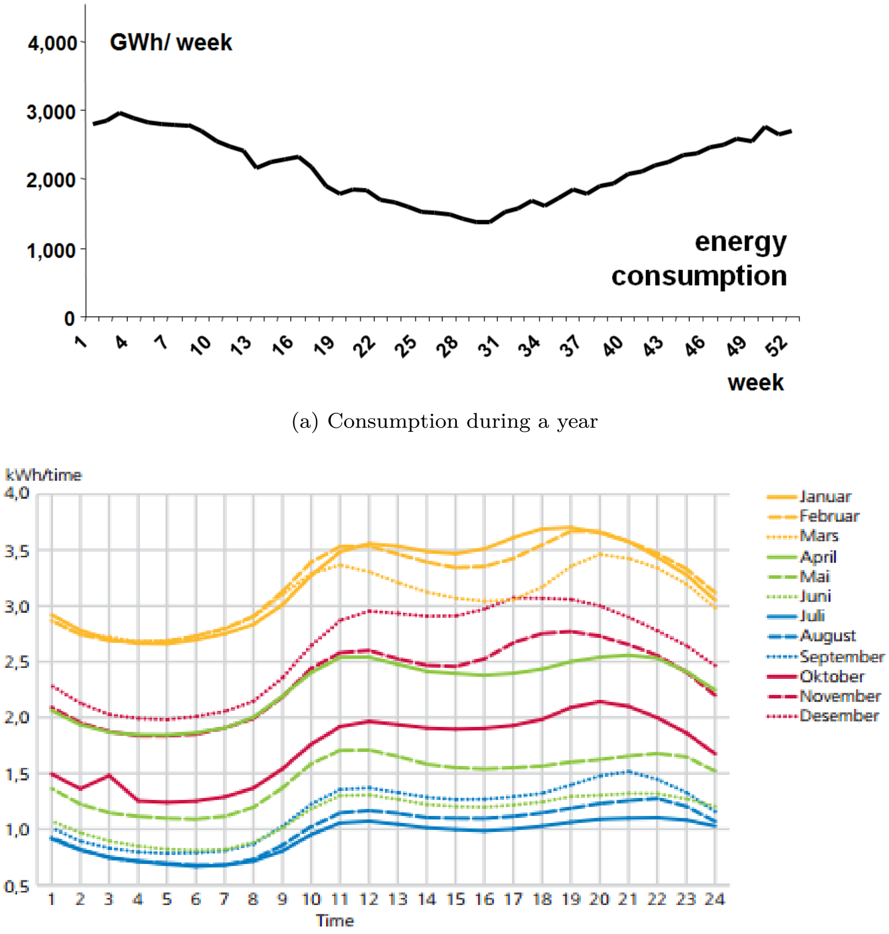
Villumsen and Philpott [33] consider the particular problem to decide on investments in electricity networks with transmission switching. They formulate the problem as a stochastic optimization problem by introducing (economic) scenarios. Every scenario ξ describes a reasonable pattern of demand and supply in the network, for which the power has to be generated at the costs $c(\cdot, \xi, \cdot)$. These costs to generate the electric power are aggregated in a single objective function by assigning probability weights to every scenario ξ. In this way Villumsen and Philpott propose the two-stage stochastic optimization problem in the form presented in (2),

$$\min_{y \in Y} \mathbb{E} \left(\min_{z \in Z(y,\xi)} c\left(y, \xi, z\right) \right). \tag{19}$$

This model balances investment costs ($y \in Y$) against potential reductions in operational cost ($z \in Z$): the inner minimization $\min_{z \in Z(y,\xi)} c(y, \xi, z)$ models the decision of the system operator to generate and supply the power, which corresponds to the actual demand ξ in the network. It is the optimal power flow problem and the decision $z \in Z$, the *wait-and-see* decision, may differ for different scenarios ξ. The expectation $\mathbb{E}$ summarizes the costs corresponding to ξ according the respective probability weight.

The *here-and-now* decision $y \in Y$ identifies the optimal network design, which is an investment decision to be established today. The investment decision consists in finding reasonable places to install automated switching devices (for example FACTS, flexible AC transmission system devices), which operate on remote basis.

It was observed that switching off existing lines may increase the overall efficiency and economic profitability of an electricity network (cf. Potluri and Hedman [27]). This fact (this paradox) is perhaps counter-intuitive, but it is clearly a starting point to incorporate switching possibilities in the network in order to adjust the power flow and to ensure constant, high profitability. This holds true even more as electricity demand, as well as its supply, are random: weather (as wind, rain or insolation), water-level of rivers and reservoirs, as well as outside temperature, time of day and time of the year are influencing factors amongst various others (Figure 1a exemplary displays the demand profile over a year for nordic countries). The expansion of renewable energy sources currently increases the respective (stochastic) volatility of electricity supply. This situation creates a window of opportunity for power network design.



(a) Consumption during a year

(b) Profiles of energy consumption in nordic countries

Figure 1: Energy consumption during a year

It was realized recently that the error caused by linearizing the inner problem in (1) exceeds the effect which is obtained by choosing different investment decisions $y \in Y$ (Fuller and Soroush [9] point this out for switching in transmission systems). A solution of the simplified, linear DC approximation is possibly a misleading candidate for the genuinely nonlinear problem (the AC formulation). Solving the linearized problem, even with high accuracy, thus does not allow justified conclusions to the real world problem. For this reason problem (1) has to be considered in its nonlinear formulation.

Classification. For the investment problem (19) the inner optimization is nonlinear, whereas the outer minimization is combinatorial. The problem thus can be classified to be of *mixed integer, nonlinear, non-convex, combinatorial stochastic optimization* type. However, the evaluation of related functions is not expensive and analytic expression for derivatives are available. The optimal power flow problem in a stochastic environment thus nicely exposes the difficulties related to the nonlinear characteristics of (1) (cf. Bukhsh et al. [6]) – a main reason why transmission switching was chosen to outline general solution techniques for stochastic optimization problems as (19).

4.1 Power flow equations

Electric power is generated at generation buses $i \in G$ (cf. the nomenclature on the next page) and transferred to the load buses to satisfy the demand there. The demand is a known quantity at every load bus in a network. The AC power flow within the network then simultaneously satisfies at every bus $i \in B$ the real power balance equations

$$\sum_{k\in B} V_i V_k \left(G_{ik}\cos\theta_{ik} + B_{ik}\sin\theta_{ik}\right) + P_i^d - P_i^g = 0 \qquad (i \in B) \tag{20}$$

and the reactive power balance equations

$$\sum_{k\in B} V_i V_k \left(G_{ik}\sin\theta_{ik} - B_{ik}\cos\theta_{ik}\right) + Q_i^d - Q_i^g = 0 \qquad (i \in B\backslash G), \tag{21}$$

which are derived from Kirchhoff's laws.

The power flow problem assumes that the real power generated P_i^g, and the voltage magnitude V_i are given quantities at generator buses (for this reason, generator buses are known as PV buses, $i \in G$). The net and reactive power demand (P_i^d and Q_i^d) are known at the load buses (PQ buses, $i \notin G$). A solution of the balance equations (20) and (21) consists of voltage angles θ_i for all buses and the voltage magnitudes V_i at the remaining buses, the load buses. Table 1 collects the known quantities and the variables of the AC power flow equations.

bus i	given quantity	variables to be determined	balance equations
generator bus (PV bus)	P_i and V_i	θ_i	(20)
load bus (PQ bus)	P_i and Q_i	V_i and θ_i	(20) and (21)

Table 1: Balance equations and variables characterizing the power flow problem, (20) and (21)

The balance equations (20)–(21) do not specify a power flow solution uniquely, as for example uniformly shifting all voltage angles θ_i of a solution by a constant angle (θ^*, say) solves the power

Nomenclature

(B, L)	graph of the transmission network consisting of buses B and transmission lines L
$i,\, k, \ldots \in B$	buses $i,\, k, \ldots$
$L \subset B \times B$	the set of transmission lines linking buses: $(i, k) \in L$ if a transmission line connects i and k
$g \in G \subset B$	generators (PV bus)
$d \in B \backslash G$	demand, or load bus (PQ bus)
V_i	voltage magnitude at bus i (measured in Volt, V)
θ_i	voltage angle (phase) at bus i (measured in radian)
$\theta_{ik} = \theta_i - \theta_k$	difference of voltage angles at buses i and j
j	imaginary unit, $j^2 = -1$
$S = P + j\, Q$	complex power (measured in Watt, W)
$P_i = P_i^g - P_i^d$	net power injected at bus i. The superscripts g indicates power generated at bus i, while d relates to demand
$Q_i = Q_i^g - Q_i^d$	net reactive power injected at bus i
$Z_{ik} = R_{ik} + j\, X_{ik}$	impedance (measured in Ohm, Ω). R_{ik} is the resistance of transmission asset linking the buses i and k, X_{ik} the reactance
$Y = \frac{1}{Z} = G + j\, B$	admittance (measured in Siemens, Ω^{-1}). G_{ik} is the conductance, B_{ik} the susceptance

flow equations equally well. To select a unique solution from the manifold of solutions the voltage phase θ of a selected generator bus (the reference, or slack bus) is occasionally fixed.

To abbreviate the notation the unknown variables are collected in a vector z, i.e., we set

$$z := \big((\theta_i)_{i \in B},\, (V_i)_{i \in B \backslash G}\big)\,.$$

We summarize the balance equations (20) and (21) as well in vectors and set

$$h(z) := \begin{pmatrix} h_P(z) \\ h_Q(z) \end{pmatrix} = \begin{pmatrix} h_P(\theta, V) \\ h_Q(\theta, V) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0. \tag{22}$$

$h_P(z)$ and $h_Q(z)$ represent equations (20) and (21), respectively. Equation (22) is simply referred to as power flow equation, or *AC power flow equation*.

An efficient and well established method to numerically solve the power flow equations (22), $h(z) = 0$, is by applying Newton–Raphson's method. The method iteratively defines a sequence z_k

by

$$z_{k+1} = z_k + \Delta z_k \quad \text{and} \quad h'(z_k) \cdot \Delta z_k = -h(z_k), \tag{23}$$

where h' is the Jacobian matrix (the derivative with respect to z). The sequence z_k converges quadratically to a solution under general conditions, provided that the starting value z_0 is already close enough to a solution of the power flow equation.

4.2 Solution techniques for the AC power flow equations

Various techniques have been considered and developed in the past to approximately, or efficiently solve the nonlinear power flow equations (22). An important approximation is based on linearization, which is natural if the phase angles in the network are almost parallel, that is, if $\theta_{ik} = \theta_i - \theta_k \approx 0$ for all directly connected buses $(i,k) \in L$.

Linearization, the DC formulation. The linearized equations derive from the observations $\cos\theta_{ik} \approx 1$ and $\sin\theta_{ik} \approx \theta_{ik}$ (the approximation are of second order for small θ_{ik}), by neglecting the conductance G_{ik} and assuming that the deviations from $V = 1$ are very small. The approximating equations obtained are

$$\begin{aligned} P_i &= \sum_k B_{ik}\left(\theta_k - \theta_i\right) \text{ and} \\ Q_i &= \sum_{k \neq i} B_{ik}\left(V_k - V_i\right) - \sum_k B_{ik} V_i, \end{aligned} \tag{24}$$

often referred to as *DC power flow equations*. In contrast to the above AC power flow equation (22), the system (24) is linear in the variables θ and V, and thus comfortably easy to solve. As already addressed, the approximation quality of the solution of (24) is typically not satisfactory. The solution of the approximating problem, however, is an important starting point for iterative methods (cf. Hedman et al. [14, 15], Hedman and Oren [13]), for example for the nonlinear solvers addressed in what follows.

The system of linear equation, which has to be solved in (23), can be computed efficiently as the derivative

$$h' = \begin{pmatrix} \frac{\partial h_P}{\partial \theta} & \frac{\partial h_P}{\partial V} \\ \frac{\partial h_Q}{\partial \theta} & \frac{\partial h_Q}{\partial V} \end{pmatrix} \tag{25}$$

in typical applications is a sparse matrix with entries

$$\frac{\partial h_{P,i}}{\partial \theta_i} = \sum_{k \neq i} V_i V_k \left(-G_{ik} \sin\theta_{ik} + B_{ik} \cos\theta_{ik}\right), \tag{26}$$

$$\frac{\partial h_{P,i}}{\partial \theta_\ell} = V_i V_\ell \left(G_{i\ell} \sin\theta_{i\ell} - B_{i\ell} \cos\theta_{i\ell}\right) \qquad (\ell \neq i), \tag{27}$$

$$\frac{\partial h_{P,i}}{\partial V_i} = 2G_{ii} V_i + \sum_{k \neq i} V_k \left(G_{ik} \cos\theta_{ik} + B_{ik} \sin\theta_{ik}\right) \text{ and} \tag{28}$$

$$\frac{\partial h_{P,i}}{\partial V_\ell} = V_i \left(G_{i\ell} \cos\theta_{i\ell} + B_{i\ell} \sin\theta_{i\ell}\right) \qquad (\ell \neq i) \tag{29}$$

in the upper row of the matrix (25) (the lower row for the derivatives of $h_{Q,i}$ being analogous). Note that if the buses i and ℓ are not connected by a transmission line, $(i, \ell) \notin L$, then the admittance is $Y_{i\ell} = G_{i\ell} + j\, B_{i\ell} = 0$ and the entries in the derivative vanish, $\frac{\partial h_{P,i}}{\partial V_\ell} = \frac{\partial h_{P,i}}{\partial \theta_\ell} = 0$. For a typical power flow network the derivative h' is a sparse, connected matrix. Its entries reflect the adjacency matrix of the graph (B, L).

The classical convergence theory for the Newton procedure requires that the derivatives are Lipschitz continuous, as outlined in Theorem 15 (cf. also Kantorovich's theorem, two different proofs are given in Kantorovich [16], Kantorovich and Akilov [17]). The following lemma uncovers that this is the case for the power flow problem, a uniform Lipschitz constant can be chosen on bounded domains.

Lemma 20. *As an operator from ℓ^∞ to ℓ^∞, the derivative h' is uniformly bounded by*

$$\|h'(y)\| \le 2\,(1 + V^{max})^2 \cdot \max_i \sum_k |Y_{ik}| = 2\,(1 + V^{max})^2 \cdot \|Y\|\,, \tag{30}$$

where Y is the bus admittance matrix and $|V_i| \le V^{max}$ for all buses $i \in B$. $\|Y\|$ is the norm of the admittance matrix, as an operator $Y : \ell^\infty \to \ell^\infty$.

Moreover it holds that h' is continuous, i.e, there is a constant c such that $\|h'(y) - h'(\tilde{y})\| \le c\,\|y - \tilde{y}\|$. As in (30) the constant c depends on $(1 + V^{max})^2$.

Proof. The norm of a matrix J induced by ℓ^∞ is

$$\sup_{\|x\|_\infty \le 1} \|J \cdot x\|_\infty = \sup_{\|x\|_\infty \le 1} \max_i \left| \sum_j J_{ij} x_j \right| = \max_i \sup_{\|x\|_\infty \le 1} \left| \sum_j J_{ij} x_j \right| = \max_i \sum_j |J_{ij}|\,.$$

Applied to the Jacobian $J = h'(z)$ with entries (26)–(27) it holds that

$$\begin{aligned} \sup_{\|x\|_\infty \le 1} \|h'(y) \cdot x\|_\infty &\le \max_i \left\{ \begin{array}{l} 2\,|Y_{ii}|\,|V_i| + \sum_{k \ne i} |V_k|\,|Y_{ik}| + |V_i| \sum_{\ell \ne i} |Y_{i\ell}| \\ + \sum_{k \ne i} |V_i|\,|V_k|\,|Y_{ik}| + |V_i| \sum_{\ell \ne i} |V_\ell|\,|Y_{i\ell}| \end{array} \right\} \\ &= \max_i \left\{ 2V^{max} \sum_k |Y_{ik}| + 2V_{max}^2 \sum_k |Y_{ik}| \right\} \le 2\,(V_{max} + 1)^2 \sum_k |Y_{ik}|\,, \end{aligned}$$

as $|G_{ik} \cos\theta_{ik} + B_{ik} \sin\theta_{ik}| \le \sqrt{G_{ik}^2 + B_{ik}^2} = |Y_{ik}|$.

Lipschitz continuity of the derivative h' itself follows by the same reasoning as above, as h'' can be provided explicitly, although collecting the terms is cumbersome. □

Gauss–Seidel. There exist further methods to solve the AC power flow equations (22). The Gauss–Seidel method is based on the observations

$$\frac{\partial h_{P,i}}{\partial \theta_i} + \sum_{\ell \ne i} \frac{\partial h_{P,i}}{\partial \theta_\ell} = 0$$

(equations (26)–(27)) and

$$\frac{\partial h_{P,i}}{\partial V_i} \gtrapprox \sum_{\ell \ne i} \frac{\partial h_{P,i}}{\partial V_\ell}$$

(equations (28)–(29)), such that the Jacobian matrix (25) is apparently a (column) diagonal dominant matrix (it is indeed a diagonal dominant matrix in the case $\theta_{ik} = \theta_i - \theta_k \approx 0$). Diagonal dominance is a simple criterion to ensure convergence of the related Gauss–Seidel method. Although the Jacobian is not diagonally dominant in general, the Gauss–Seidel method often converges in practice and provides reasonable solutions.

The method to solve the power flow equations (22) requires less memory than the Newton–Raphson method, although it is typically slower.

Fast decoupled method. The fast decoupled method takes advantage of the fact that practical power transmission lines have a high $^X/_R$-ratio (i.e., $G \ll B$ and G can be neglected), and in this situation the Jacobian (25) can be approximated by

$$h' \cong \begin{pmatrix} \frac{\partial h_P}{\partial \theta} & 0 \\ 0 & \frac{\partial h_Q}{\partial V} \end{pmatrix}. \tag{31}$$

Substituting the initial equations (20) and (21) in the simplified Jacobian (31) it turns out that h' is a constant matrix. The matrix thus has to be inverted just once at the beginning of the iterations.

Rectangular coordinates. We finally mention that the AC power flow equations (22) can be formulated in rectangular coordinates as well. The solution techniques change then accordingly. The before mentioned overview Bienstock [2] covers this aspect as well.

4.3 The optimal power flow problem – economic dispatch

The optimal power flow problem is the economic problem of minimizing the total production costs in the transmission network, while the given demand has to be met and satisfied. In order to reduce costs, the system operator may adjust the power production of different generator units. The optimal power flow problem is often stated in the form

$$\underset{\text{in } V, \theta, P \text{ and } Q}{\text{minimize}} \quad \sum_{g \in G} c_{g,P} \cdot P_g + c_{g,Q} \cdot Q_g \tag{32}$$

$$\text{subject to } h(V, \theta;\, P, Q) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{33}$$

$$|S_{m,n}| \le S^{max},\ |S_{n,m}| \le S^{max} \tag{34}$$

$$V_i^{min} \le V_i \le V_i^{max} \qquad (i \in B), \tag{35}$$

$$\Delta\theta^{min} \le \theta_i - \theta_k \le \Delta\theta^{max} \qquad (i \in B), \tag{36}$$

$$P_g^{min} \le P_g \le P_g^{max} \qquad (g \in G), \tag{37}$$

$$Q_g^{min} \le Q_g \le Q_g^{max} \qquad (g \in G). \tag{38}$$

The nonlinear power flow equations are discussed in detail in the previous section, cf. (22). The line flow constraints (34) on the complex power $S_{n,m}$ involve real and reactive power on the line (n, m), which are the individual components (summands) in the sums (20) and (21) (cf. Sahraei-Ardakani et al. [30]).

These equations enter the *optimal* power flow problem as nonlinear equality constraints in (33). The power flow constraints are denoted $h(V, \theta;\, P, Q) = 0$ to express that P_g and Q_g, the net and reactive power injected at the generators $g \in G$, are additional variables, which are free variables in the optimal power flow problem. They represent the real and reactive power, which can be regulated by the system operator in order to minimize the total production costs with the restriction to satisfy the demand.

The constraints (35) and (36) insure that the voltage level is met in all buses within a limited bandwidth and difference of phase angles, while the constraints (37) and (38) control the power, and net power injected at the generator buses. These constraints (35)–(38) represent box constraints.

In order to write the problem concisely it is comfortable to aggregate the variables again as

$$z := \big((V_i)_{i\in G}, (\theta_i)_{i\in B}, (P_i + jQ_i)_{i\in G}\big)\,. \tag{39}$$

The comprised form of the optimal power flow problem (32)–(38) thus is

$$\begin{aligned} \text{minimize } & c(z) \\ \text{subject to } & h(z) = 0, \\ & z \in \mathrm{B}, \end{aligned}$$

where c is a general cost function. For ease of presentation we include the line limits (34) in the box constraints (34)–(38) and rewrite them as $z \in \mathrm{B}$ (cf. Footnote 2 on page 9).

5 Outline of the test problem

The two-stage stochastic optimization problem we address here as a test case is taken from transmission switching and formulated as an optimal investment problem. We consider a transmission network under different demand loads (scenarios) ξ. These load scenarios represent future demands of the network considered. Based on these loads it is to be decided if it is beneficial to install an automated transmission switch, and where this switch should be located. Villumsen and Philpott [33], from which the following statement is cited, describe the problem in further detail. "*Note, that even though the fixed cost of enabling a line to be switched instantaneously may be small (e.g., if the switch is already present and only communication equipment needs to be installed) it may not be worthwhile to enable switching on all lines (unless this cost is 0 for all lines), since some lines may never be switched.*"

Objective and cost function. Installing a switch at the line $y \in L$ (this is the outer problem) alters the network. The switch can be leveraged (inner problem) after scenario ξ has materialized. Once the demand $S_i(\xi)$ is realized at all nodes $i \in B$, the inner problem assigns the power generators with the objective to produce the energy as cheaply as possible. The inner problem thus consists in solving the optimal power flow problem (32)–(38) for every scenario ξ.

The line $y \in L$, where the two-stage stochastic optimization problem

$$\min_{y\in L} \mathbb{E}_\xi \left(\min_{z\in Z(y)} c\,(y, \xi, z) \right) \tag{40}$$

attains its minimum, has the highest cost savings on average for all the scenarios considered. This is the line where the new switch should be installed and solving (40) identifies this line.

The cost function of the inner problem, as stated in (32), collects injected real and reactive power in a linear way,

$$c(y, \xi, z) = \sum_{g\in G} c_{g,P}\, P_g + c_{g,Q}\, Q_g.$$

Notice that all components of z (cf. (39)) are random, i.e., $z = z(\xi)$: the components $P_i(\xi)$ and $Q_i(\xi)$ are explicit under scenario ξ, while $V_i(\xi)$ and $\theta_i(\xi)$ result from the inner optimization.

The optimization problem (40) is discrete, and it can be solved, in principle, by inspecting the objective for each $y \in L$ separately. Because of long computation times this strategy can only be considered for small networks. The related time amount explodes even, if not just one line is subject to switching, but two or even more lines, $(y_1, \dots y_n) \in L^n$. The combinatorial problem of selecting the best combination of transmission lines cannot be solved by inspection and different solution techniques thus have to be considered. For this we address a heuristic in Section 5.2 below.

5.1 Test cases for the discrete problem

The transmission networks for the test cases we consider are contained in matpower (cf. Zimmerman and Murillo-Sánchez [36], Zimmerman et al. [37]) and include test cases from Poland with a number of buses varying from 2 383 to 3 375 (see Table 2 for some characteristics). They represent winter and summer, as well as morning and evening peak loads between 1999 and 2008. As well we have run the IEEE 118[5] bus network, which is more frequently addressed in the literature. We have chosen to present this test case here, as the results show similar patterns for many other grids.

Load scenarios, scenario generation. The literature follows several approaches to determine future scenarios. These approaches often can be distinguished between applications having long term, or short term horizons in mind. The scenarios in long term horizons typically reflect expert opinions on future developments. Villumsen et al. [34], for example, follow this approach by choosing different future scenarios of the economy, each describing likely developments of the society twenty years ahead in time. The problem formulation (40) applies for both.

Our application is short term. It tries to identify specific transmission lines, where profits may be expected if a switch device is installed such that the line can be switched off during some hours during the day at this point.

To this end we base our scenarios on real load patterns (real and reactive power P_i, Q_i at the buses $i \in B$), which reflect the power at some instants of time during a day and during a year. By taking a snapshot on hourly basis during a year a sample of size $24 * 365$ is obtained, the dimension of each snapshot is given by the number of buses in the network (the dimension is 118 for the IEEE 118 test case, e.g.). The loads observed at all buses follow a multivariate empirical distribution.

In this case study we simulate these scenarios by employing multivariate log-normal distributions,[6] which are based on the following parameters:

(i) the mean of the observations follows the average load displayed in Figure 1b;

(ii) a (time depending) variance is employed at every individual bus to reflect its varying consumption pattern over time at this bus. The parameter is extracted from data provided by the British National Grid (cf. Electricity Ten Year Statement), and finally

(iii) a covariance is imposed between the buses via a covariance matrix. A high correlation coefficient accounts for the fact that different buses show a similar consumption behavior. This is indeed the case, as individual households switch on their light, TV, the heating or air conditioner, e.g., at about the same time during a day. Energy consumers at different buses act in a co-monotone way and their loads thus are highly correlated.

[5] cf. Matpower or the webpage http://www.ee.washington.edu/research/pstca/

[6] This modification of the normal distribution ensures nonnegative outcomes.

network		IEEE 118	2383wp	2746wp	2736sp	3120sp
characteristics	dimension (i.e., buses)	118	2 383	2 746	2 736	3 120
	lines	186	2 896	3 514	3 504	3 693
representative scenarios		distance d_1 of the reduced measure				
100		1.8	4.3	7.8	6.2	3.7
10		2.4	5.2	9.3	7.7	4.7
7		2.6	5.6	9.9	8.4	5.2
5		2.8	6.1	10.7	9.0	5.6

Table 2: Wasserstein distance d_1 of the initial 8 760 (= 24 * 365) scenarios and the reduced distribution, located on 100, 10, 7 and 5 representative scenarios. The results are displayed for the IEEE 118 and for 4 different Polish test cases (winter peak, summer peak).

Scenario reduction. Theorem 5 justifies employing different probability measures to evaluate the objective. Further, Section 2.3 outlines that clustering methods can be employed to reduce the number of scenarios to be considered. We combine these results to reduce the number of scenarios in total to a reasonably small amount, which can be treated in further computations (cf. also Pflug and Pichler [22, 23]).

Table 2 collects the Wasserstein distance of the original sample, compared to a reduced (clustered) sample by employing the weighted ℓ^2-norm (weighted Euclidean norm, $\|x\|^2 = \frac{1}{m}\sum_{i=1}^{m} x_i^2$). The table displays the results for 5, 7, 10 and 100 representative scenarios.

It is apparent from Table 2 that approximations with more representative points are better, they are closer to the original distribution. However, to improve the approximation quality considerably, significantly more representative points have to be accepted in the approximating measure. Increasing the number of representative points from 5 to 100, for example, improves the precision by less than 50 % in all examples outlined in Table 2 (cf. Remark 3, the curse of dimensionality).

Computational results. The test case IEEE 118 has 186 lines. In its genuine setting provided by matpower, 24 of these lines can be switched off by obtaining cost savings. The best line (line 104) leads to savings of about 0.4 ‰. Table 3a relates the behavior of the network IEEE 118 in this genuine setting with real distributions (indicated are the standard deviation and the correlation of the different distributions used to run (40)). The table shows a rather small dependence on the correlation ρ. Half of the lines, which lead to savings in the genuine scenario, still lead to savings under much higher standard deviation.

The situation does not differ significantly for the Polish network 2 383, winter peak (Table 3b). A smaller correlation and higher standard deviation reduce the number of lines, which lead to savings in the stochastic case as well.

We conclude from these tables that being a savings line is a stable property. Once a savings line is identified, then with high probability the line will provide savings in a stochastic environment as well, even if the standard deviation of the future distribution is high.

correlation	ρ:	90 %	50 %	0 %
volatility σ:	10 %	22	22	22
	25 %	15	14	13
	50 %	11	10	13

(a) Test case 118: number of lines out of the original savings lines, which lead to cost savings in the stochastic case

	ρ:	90 %	50 %	0 %
σ:	10 %	10	10	9
	25 %	9	10	9
	50 %	10	5	3

(b) Test case 2383 winter peak: number of lines out of the 10 best savings lines, which lead to cost savings in the stochastic case

Table 3: Savings lines under different distributions

5.2 Continuous switching, the relaxed problem

Switching off a transmission line $(i,k) \in L$ corresponds to setting the corresponding admittance to zero, $Y_{i,k} = 0$. It is hence possible to switch off the transmission line (i,k) by continuously *sliding* the admittance from $Y_{i,k}$ (the line in full, 100 % service) down to 0 (the line being completely switched off). In this way the combinatorial problem (40) of finding transmission lines to be switched off can be studied by considering the continuous relaxation instead.

Fuller et al. [10] elaborate a heuristic to identify transmission lines, which offer a potential for savings when being switched off. The heuristic they propose is related to continuous switching, as it is based on ranking the dual variables (shadow prices) associated with transmission lines in full service. Their procedure (in a nutshell) selects those lines, which have a positive shadow price in full service and ranks them accordingly. In Figures 2a and 2b, a positive shadow price corresponds to a positive slope for the line in full (100 %) service.

Figure 2a displays production costs for the IEEE 118 bus test case, where every single transmission line is continuously faded out. Every line in the plot corresponds to a transmission line:

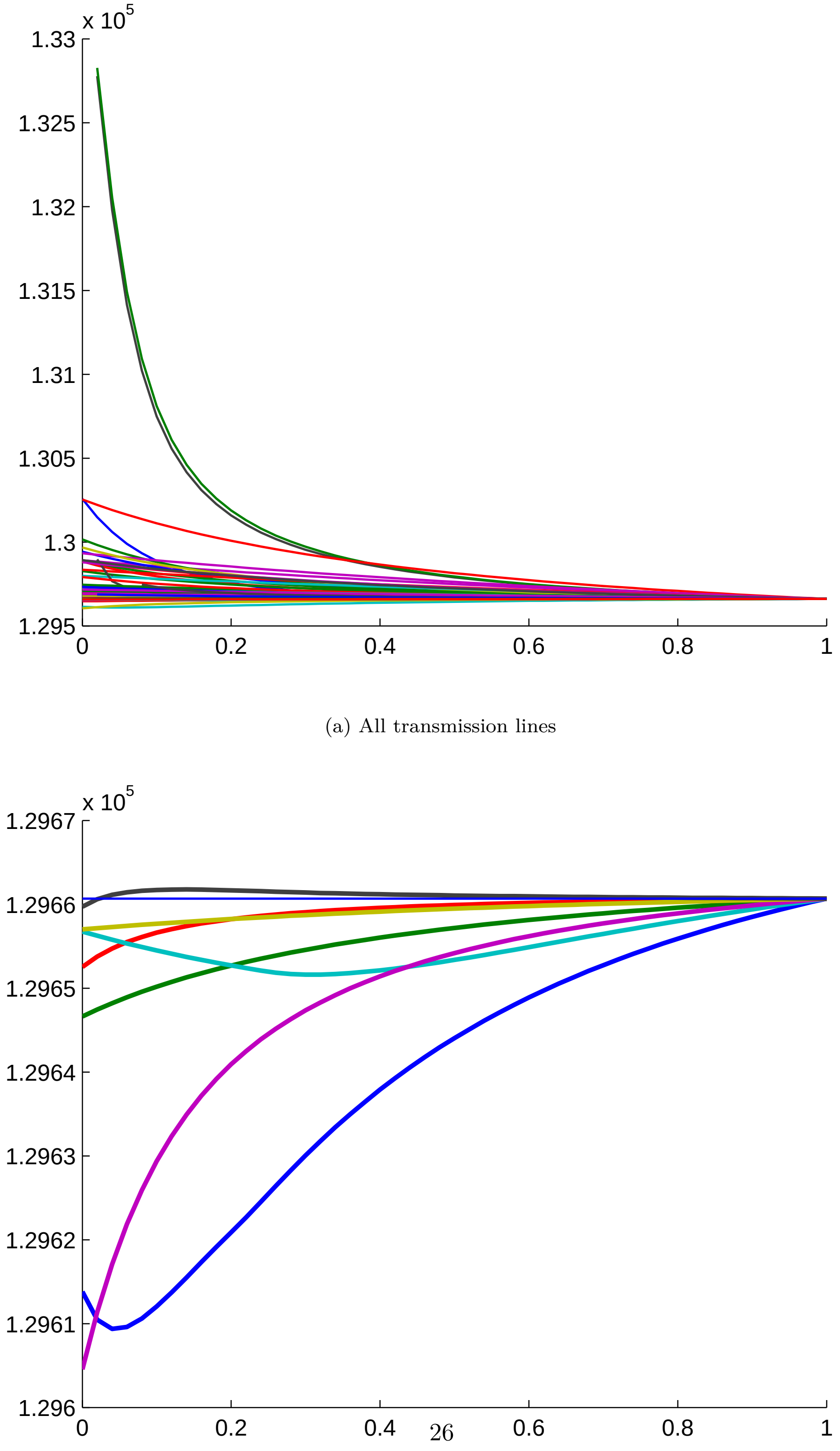
(i) Production costs are apparently unbounded for two transmission lines. These two lines cannot be switched off, as some demand buses would be cut off from energy supply.

(ii) Decreasing the admittance to 0 corresponds to increasing the impedance or the resistance. It is not to be expected that the graph corresponding to a transmission line is defined for every admittance between 0 (out of service) and 1 (full service), as this impacts feasibility. This is, however, the case for all lines in the IEEE 118 test case (Figure 2a).

Figure 2b collects production costs for varying admittance of 7 selected transmission lines. These are the lines which show savings for reduced admittance. The total savings potential, however, is small, it is within a range up to 0.4 ‰ for the IEEE 118 test case. We have observed a total savings potential of less than 1 ‰ in many other test cases as well.

5.3 What-if analysis, and discussion of the test problem

Economic savings of the transmission switching problem are marginal in usual situations, the cost saving we observe are often less than 1 ‰ of the total production costs. The savings we observe are



(a) All transmission lines

(b) 7 selected lines

Figure 2: IEEE 118 bus test case. Total production costs by sliding the admittance of an individual transmission line between 0 % (left, line out of service) and 100 % (right, line in full service)

higher in (highly) congested networks (indeed, congestion management has been the main reason for line switching in the previous literature).

The Figures 3–5 expose continuous switching in different configurations. We display the results in a what-if analysis for 7 individual, representative scenarios obtained by scenario reduction, as outlined in Section 2.3. It is a repeated pattern that the curves corresponding to continuous switching are almost parallel for the different scenarios. We interpret this behavior by saying that the heuristic is a reliable indicator to predict savings lines, irrespective of the individual load.

Figure 3 displays continuous switching for the baseline scenario (thick, this is the load taken from the matpower file) and under 7 scenarios after reducing the distribution as described in Section 5.1. While savings can be expected on the line 32 (Figure 3a), production costs increase for the line 105 (Figure 3b).

Figure 4 addresses line 32 (the 2nd-best individual line to achieve savings), which is already depicted in Figure 3a. Displayed are now the results for *independently* chosen scenarios (the correlation coefficient is $\rho = 0$) for varying variance ($\sigma = 10\,\%$ in Figure 4a, and and $\sigma = 0.1\,‰$ in Figure 4b).

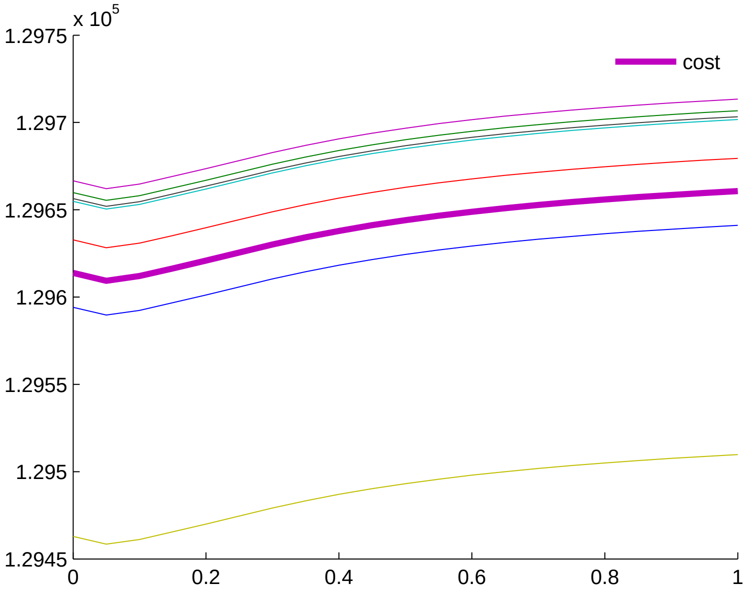
Figure 5a finally shows the impact of correlation. As mentioned, a high correlation among nodes in the electricity network is realistic, as electricity demand in private households is of course correlated (lights are switched on in the evening, heating units are switched on in case it is cold, and air conditions are activated if it is hot).

All plots in Figures 3–5 show a common pattern. Irrespective of the correlation or the variance, and irrespective of the particular line, switching off a transmission line exposes a pattern, which is a characteristic of the transmission line itself. The particular scenario or load pattern in the network is not important, the important driving factor is the particular line in the network topology. This is the case for individually drawn scenarios, as well as for representative scenarios obtained by clustering according the Wasserstein distance as outlined in Section 2.

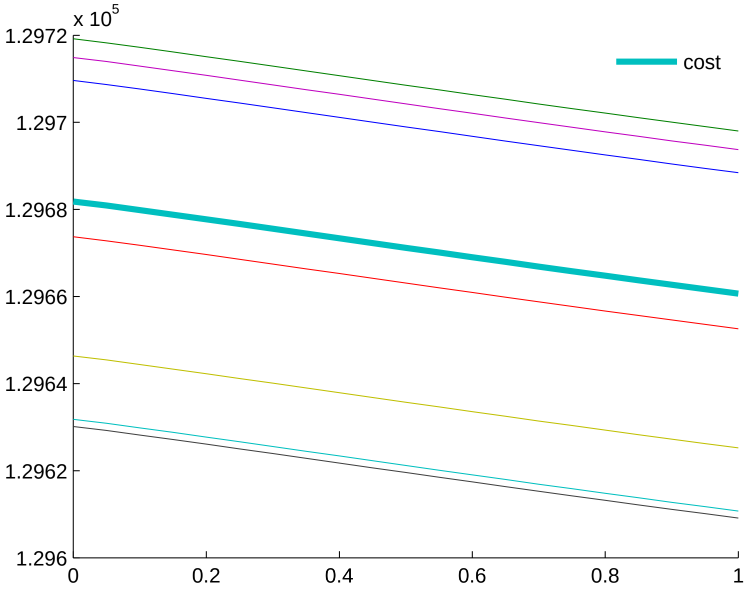
Algorithm 1 and the scenario reduction based on Wasserstein balance each other. It is desirable to increase the number of scenarios to obtain a higher precision. We have seen in Table 2, however, that considerable improvements may only be obtained by choosing significantly more representative scenarios, the relation of the corresponding Wasserstein distance and the number of scenarios is extremely *disproportionate*. The curse of dimensionality (cf. Remark 3) imposes very restrictive limitations on increasing the number of scenarios, as the dimension of the problem is huge (see Table 2). More scenarios explode computation times, but Algorithm 1 is designed to accelerate the computation for similar scenarios.

The predictor in Algorithm 1 (see the step (iii)) does not help much for a small number of scenarios, as these scenarios are not similar in this case. The predictor reduces total computation time by a factor of approximately 5 if many scenarios are similar. In other words, the algorithm can handle 5 times more recourse problems (scenarios) in about the same time, compared to a usual Newton procedure. Parallelization can be used as well to reduce the total computation time of the decomposed problem (8), but again by not more than a factor, the number of parallel processors.

Regarding complexity of the problem we mention as well that the algorithm is (at least) of order $\mathcal{O}(n^3)$, as $n \times n$-matrices have to be inverted. In the case study, n is the number of buses in the network. Larger networks thus can handle only $n^{-1/3}$ scenarios in about the same time, which makes clustering an essential tool, and the explicit bound in Theorem 5 a valuable error bound.

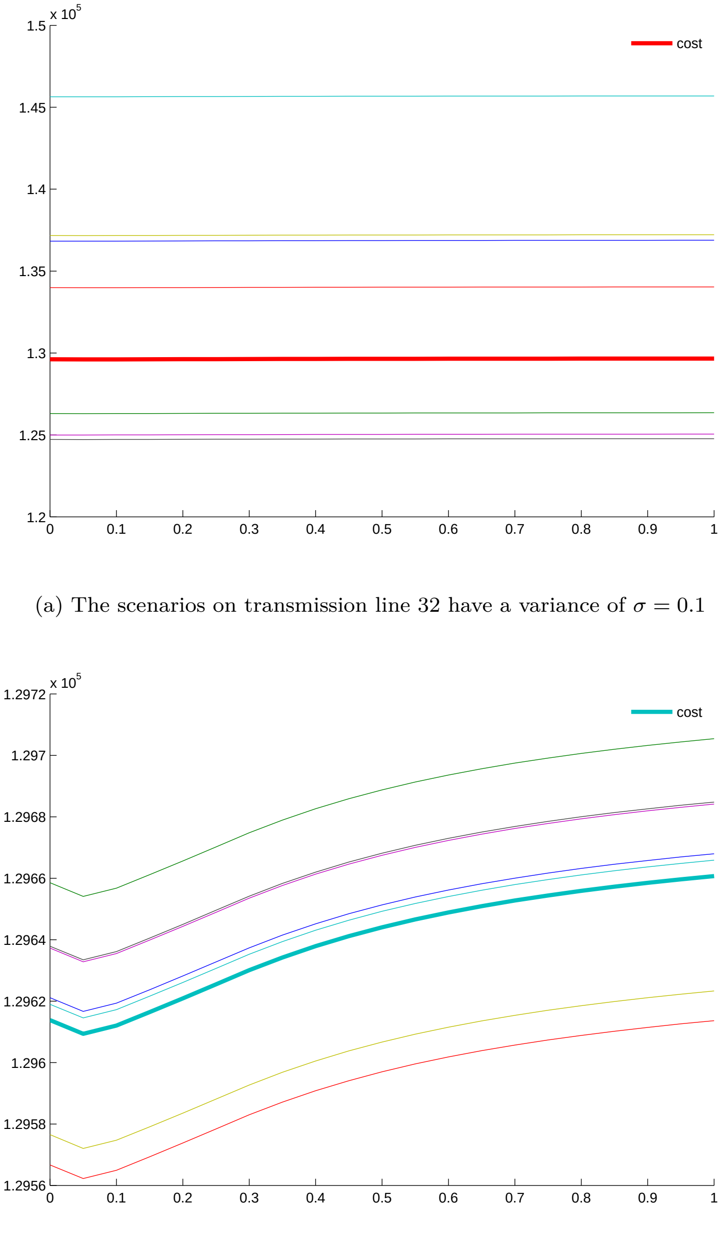


(a) Line 32: 0.4 ‰ cost savings when switched off



(b) Line 105: 0.15 ‰ higher energy production costs if switched off

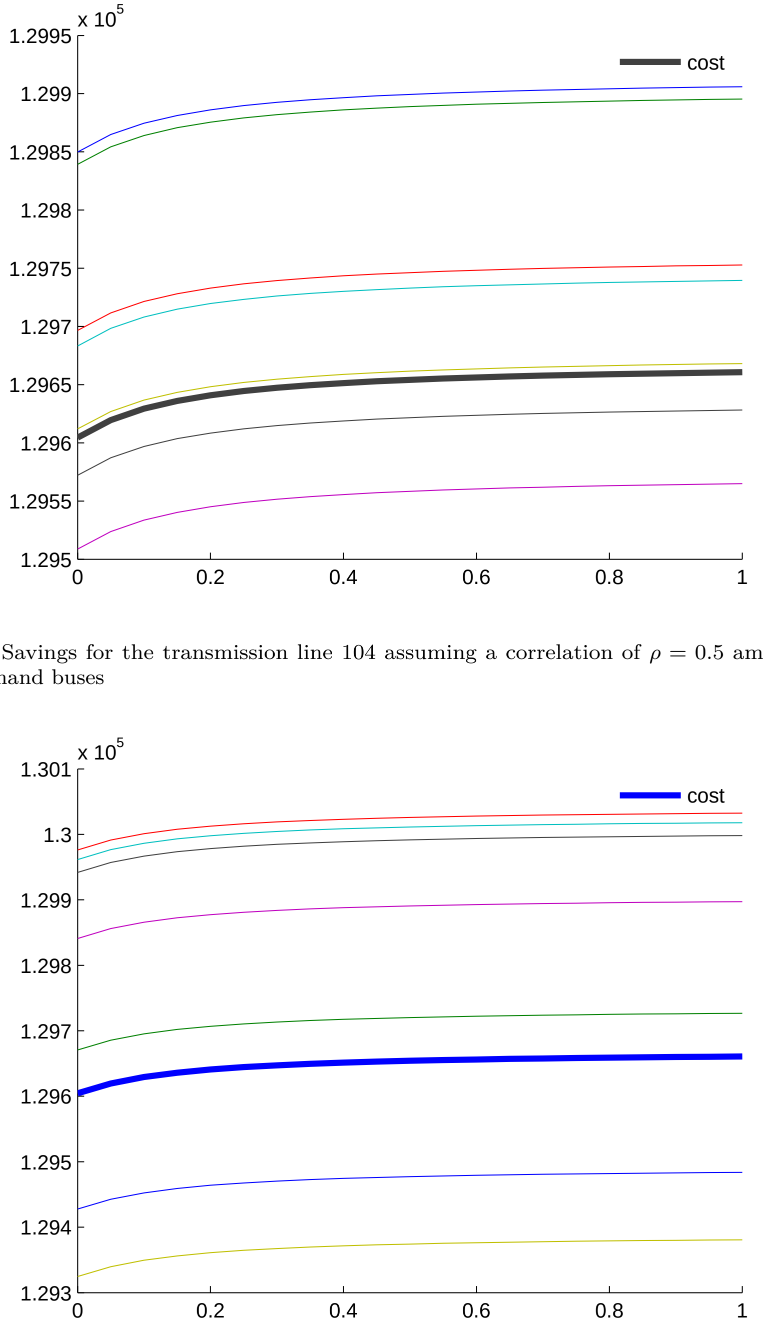
Figure 3: Continuous switching for individual transmission lines (IEEE 118) under 7 different scenarios (the thick line represents the load contained in matpower)



(a) The scenarios on transmission line 32 have a variance of $\sigma = 0.1$

(b) The scenarios on transmission line 32 have a variance of $\sigma = 10^{-5}$

Figure 4: Continuous switching for different scenarios (7 scenarios in each plot; the thick line represents the load contained in matpower)



(a) Savings for the transmission line 104 assuming a correlation of $\rho = 0.5$ among all demand buses

(b) Savings for the transmission line 104 assuming a correlation of $\rho = 1.0$ among all demand buses

Figure 5: Continuous switching for different scenarios (7 scenarios in each plot; the thick line represents the load contained in matpower)

6 Summary

6.1 Is transmission switching beneficial from economic perspective?

A positive answer to this question (or claim) was actually the starting point of our research, as it is supported by various publications. However, these publications typically consider a DC approximation, just a single scenario, or highly congested networks.

Our results on nonlinear networks in a usual operation mode do not confirm high savings, the savings we observe are typically in a marginal range up to 1 ‰ of total production costs. This range is found by a *what-if* analysis (i.e., by inspecting individual scenarios directly), as well as for the combined, full stochastic optimization problem (1).

6.2 Stability of nonlinear, two-stage stochastic optimization problems

The main result of this paper is the result on stability of nonlinear stochastic optimization programs, Theorem 5. This theorem provides an explicit upper bound for discrete approximations of probability distributions in the context of two-stage stochastic optimization programs. The bound is sharp and attained, and further computationally available. For discrete distributions, such as the empirical distribution, the result justifies clustering methods by providing an explicit error bound of related approximations.

Acknowledgment

The authors are indebted to Prof. Shmuel S. Oren (UC Berkeley) for several discussions and assessing an initial version of this paper. We wish to thank two anonymous referees for their commitment to carefully reading the manuscript and providing very serious and sound suggestions to improve the paper. We have adopted some of their formulations.

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures.* Birkhäuser Verlag AG, Basel, Switzerland, 2nd edition, 2005. doi:10.1007/978-3-7643-8722-8. 6

[2] D. Bienstock. Progress on solving power flow problems. In *Optima 93*. 2013. 14, 15, 21

[3] F. Bolley, A. Guillin, and C. Villani. Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probability Theory and Related Fields*, 137(3-4):541–593, 2007. ISSN 0178-8051. doi:10.1007/s00440-006-0004-7. 4

[4] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004. ISBN 0-521-83378-7. URL http://www.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf. 10

[5] C. G. Broyden and M. T. Vespucci. *Krylov solvers for linear algebraic systems.* Number 11 in Studies in Computational Mathematics. Elsevier, 2004. URL http://books.google.com/books?id=WoyUPP1Ps2QC. 13

[6] W. A. Bukhsh, A. Grothey, K. I. M. McKinnon, and P. A. Trodden. Local solutions of optimal power flow. *IEEE Transactions on Power Systems*, 28(4):4780–4788, Nov 2013. doi:10.1109/TPWRS.2013.2274577. 17

[7] G. B. Dantzig. Linear programming under uncertainty. *Management Science*, 1(3 and 4): 197–206, 1955. 1

[8] J. E. Dennis, Jr. On Newton-like methods. *Numerische Mathematik*, 11:324–330, 1968. 13

[9] J. D. Fuller and M. Soroush. Accuracies of optimal transmission switching heuristics based on DCOPF and ACOPF. *IEEE Transactions on Power Systems*, 2013. doi:10.1109/TPWRS.2013.2283542. 17

[10] J. D. Fuller, R. Ramasra, and A. Cha. Fast heuristics for transmission-line switching. *IEEE Transactions on Power Systems*, 27(3):1377–1386, Aug 2012. ISSN 0885-8950. doi:10.1109/TPWRS.2012.2186155. 25

[11] D. M. Gay. Some convergence properties of broyden's method. *SIAM J. Numerical Analysis*, 16(4):623–630, August 1979. URL `http://www.jstor.org/stable/2156533`. 14

[12] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag Berlin Heidelberg, 2000. doi:10.1007/BFb0103945. 4

[13] K. W. Hedman and S. S. Oren. Improving economic dispatch through transmission switching: New opportunities for a smart grid, 2009. 19

[14] K. W. Hedman, R. P. O'Neill, E. B. Fisher, and S. S. Oren. Optimal transmission switching – sensitivity analysis and extensions. *IEEE Transactions on Power Systems*, 23(3):1469–1479, 2008. doi:10.1109/TPWRS.2008.926411. 19

[15] K. W. Hedman, R. P. O'Neill, E. B. Fisher, and S. S. Oren. Optimal transmission switching with contingency analysis. *IEEE Transactions on Power Systems*, 2008. doi:10.1109/TPWRS.2009.2020530. 19

[16] L. Kantorovich. Functional analysis in applied mathematics. *Uspekhi Mat. Nauk.*, 3:89–185, 1948. 20

[17] L. Kantorovich and G. P. Akilov. Functional analysis in normed spaces. 1959. 20

[18] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009. doi:10.1137/070704277. 2

[19] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980. doi:10.1090/S0025-5718-1980-0572855-7. 14

[20] G. Ch. Pflug. On distortion functionals. *Statistics and Risk Modeling (formerly: Statistics and Decisions)*, 24:45–60, 2006. doi:10.1524/stnd.2006.24.1.45. 5

[21] G. Ch. Pflug and A. Pichler. Approximations for probability distributions and stochastic optimization problems. In M. Bertocchi, G. Consigli, and M. A. H. Dempster, editors, *Stochastic Optimization Methods in Finance and Energy*, volume 163 of *International Series in Operations Research & Management Science*, chapter 15, pages 343–387. Springer, New York, 2011. ISBN 978-1-4419-9586-5. doi:10.1007/978-1-4419-9586-5. 7

[22] G. Ch. Pflug and A. Pichler. *Multistage Stochastic Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2014. doi:10.1007/978-3-319-08843-3. URL `https://books.google.com/books?id=q_VWBQAAQBAJ`. 7, 24

[23] G. Ch. Pflug and A. Pichler. Dynamic generation of scenario trees. *Computational Optimization and Applications*, 62(3):641–668, 2015. doi:10.1007/s10589-015-9758-0. 7, 24

[24] G. Ch. Pflug and W. Römisch. *Modeling, Measuring and Managing Risk*. World Scientific, River Edge, NJ, 2007. doi:10.1142/9789812708724. 5

[25] A. Pichler. The natural Banach space for version independent risk measures. *Insurance: Mathematics and Economics*, 53(2):405–415, 2013. doi:10.1016/j.insmatheco.2013.07.005. 5

[26] A. Pichler. Premiums and reserves, adjusted by distortions. *Scandinavian Actuarial Journal*, 2015(4):332–351, sep 2013. doi:10.1080/03461238.2013.830228. 5

[27] T. Potluri and K. W. Hedman. Impacts of topology control on the ACOPF. In *Power and Energy Society General Meeting, 2012 IEEE*, pages 1–7, 2012. doi:10.1109/PESGM.2012.6345676. 15

[28] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems Vol. I: Theory, Vol. II: Applications*, volume XXV of *Probability and its applications*. Springer, New York, 1998. doi:10.1007/b98893. 3

[29] A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, 2006. URL `http://books.google.com/books?id=ltHkMoxqv68C`. 9, 10

[30] M. Sahraei-Ardakani, A. Korad, K. W. Hedman, P. Lipka, and S. Oren. Performance of AC and DC based transmission switching heuristics on a large-scale Polish system. In *PES General Meeting/ Conference Exposition, 2014 IEEE*, pages 1–5, July 2014. doi:10.1109/PESGM.2014.6939776. 21

[31] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming*. MOS-SIAM Series on Optimization. SIAM, 2009. doi:10.1137/1.9780898718751. 1, 2

[32] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003. ISBN 0-821-83312-X. URL `http://books.google.com/books?id=GqRXYFxe0l0C`. 3, 4

[33] J. C. Villumsen and A. B. Philpott. Investment in electricity networks with transmission switching. *European Journal of Operational Research*, 222:377–385, 2012. doi:10.1016/j.ejor.2012.05.002. 15, 22

[34] J. C. Villumsen, G. Brønmo, and A. B. Philpott. Line capacity expansion and transmission switching in power systems with large-scale wind power. *IEEE Transactions on Power Systems*, 2012. doi:10.1109/TPWRS.2012.2224143. 15, 23

[35] T. Yamamoto. Historical developments in convergence analysis for Newton's and Newton-like methods. *Journal of Computational and Applied Mathematics*, 124:1–23, 2000. doi:10.1016/S0377-0427(00)00417-9. 13

[36] R. D. Zimmerman and C. E. Murillo-Sánchez. *Matpower 4.1.* Power Systems Engineering Research Center. URL `http://www.pserc.cornell.edu/matpower/`. 23

[37] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas. MATPOWER: Steady-state operations, planning and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19, Feb. 2011. URL `http://www.pserc.cornell.edu/matpower/manual.pdf`. 23

A Appendix

The following lemma provides that Hölder continuity is closed under minimization and maximization. It is the essential tool to provide continuity of risk functionals of the general form (4).

Lemma 21. *Let $(f_\iota)_{\iota\in I}$ be a family of Hölder continuous functions with exponent $\beta \le 1$ and constant H_β. Then the functions $\inf_{\iota\in I} f_\iota$ and $\sup_{\iota\in I} f_\iota$ are Hölder continuous as well with the same exponent β and constant H_β (provided, that $\inf_{\iota\in I} f_\iota(x_0) > -\infty$ for some x_0).*

Proof. Let $x \in X$ be fixed. For $\varepsilon > 0$ find $\iota \in I$ such that $f_\iota(x) \le \inf_{\iota\in I} f_\iota(x) + \varepsilon$. Then

$$f_\iota(y) - \inf_{\iota\in I} f_\iota(x) \le f_\iota(y) - f_\iota(x) + \varepsilon \le H_\beta \cdot d(x,y)^\beta + \varepsilon.$$

It follows thus that

$$\inf_{\iota\in I} f_\iota(y) - \inf_{\iota\in I} f_\iota(x) \le H_\beta \cdot d(x,y)^\beta + \varepsilon.$$

As $\varepsilon > 0$ is chosen arbitrarily hence

$$\inf_{\iota\in I} f_\iota(y) - \inf_{\iota\in I} f_\iota(x) \le H_\beta \cdot d(x,y)^\beta.$$

Interchanging the roles of x and y reveals that $\inf_{\iota\in I} f_\iota(\cdot)$ is Hölder continuous with the same constant H_β, for the same exponent β.

The assertion follows for the supremum as $-f_\iota$ is Hölder continuous with the same exponent, and as $\sup_{\iota\in I} f_\iota = -\inf_{\iota\in I} -f_\iota$.

□