COMPOUNDING ENERGY LTD

Working Paper Series • CE-WP-2026-03

Stabilised decomposition with embedded N-1 contingency cuts for security-constrained nodal capacity expansion

The single-level foundation of the CENovaSage AFN-B engine

Dr Christopher T. M. Clack
Founder, Compounding Energy Ltd
christopher@compoundingenergy.com

September 2026

Abstract. Investment-grade capacity expansion modelling for liberalised electricity systems must reconcile nodal resolution, transmission security, and stochastic uncertainty within a decomposition that admits valid global bounds at every iterate. This paper specifies and verifies the *single-level foundation* of the Compounding Energy CENovaSage capacity-expansion engine — the degenerate one-block case of the production three-level Adaptive-Fidelity Nested Benders (AFN-B) algorithm — and isolates two questions an engineering team must settle before scaling: how to stabilise the cutting-plane master, and how to embed N-1 security without enumerating every contingency in the master. We formulate the two-stage security-constrained problem over a full DC power-flow network, derive the LP-duality optimality cut and its validity as a global

minorant under LP-relaxed operations, and study three cut/master configurations: naive (single-dual) cuts, Magnanti–Wong Pareto cuts at an interior core point, and a level-bundle quadratic master with the lower bound read from the unregularised cut model (the bound-trap discipline). N-1 security is handled by a PTDF/LODF separation oracle that emits contingency cuts only when a monitored corridor binds. We restate the convergence results at the rigour an expert reader expects — finite termination of the mixed-integer master by a finitely-many-dual-bases argument, a total-cut bound for the oracle, and the qualitative level-bundle rate with its strong-convexity hypothesis flagged as non-binding for piecewise-linear recourse — and demonstrate end-to-end correctness on a stylised GB 30-zone, 8-scenario instance engineered so that an N-1 boundary bottleneck genuinely binds. All six configurations agree on the optimum (£74.67M) to within the 10^{-5} convergence tolerance — the naive and Magnanti–Wong configurations to machine precision, the level-bundle configuration to a relative spread of 8×10^{-6} — the power-systems and optimality invariants pass to numerical tolerance with every stored cut audited as a valid, tight minorant, and the oracle reproduces the full-enumeration optimum exactly. This is a correctness and mechanics demonstrator; performance at production scale, and the screening economy of the oracle, are the subject of the nested-engine companion work (in preparation).

Keywords: capacity expansion, transmission expansion planning, Benders decomposition, level-bundle stabilisation, Magnanti–Wong cuts, N-1 contingency, security-constrained dispatch, DC optimal power flow

Contents

1 Introduction

Investment-grade capacity expansion (CEM) of an electricity system requires the simultaneous handling of three sources of difficulty that classical commercial tools treat in isolation. First, *nodal resolution*: investment decisions for transmission, generation, and storage interact through Kirchhoff's laws and binding intra-zonal flow limits, and ignoring those interactions produces investment plans that are infeasible the moment the system is dispatched at high resolution. Second, *stochastic uncertainty*: a 15-year capital-allocation decision must hedge against the realisations of weather-driven generation, demand growth, fuel-price shocks, and policy paths, none of which are reducible to a single representative scenario. Third, *N-1 security*: the regulatory and engineering constraint that the system must remain feasible after any single transmission element trips is binding for a substantial fraction of the operational year, particularly in the GB system, where corridor congestion is the dominant source of redispatch cost.

Conventional approaches, as commonly deployed, handle one of these dimensions while neglecting the others. Aggregated zonal CEM tools (PLEXOS Long-Term, Aurora) handle stochastic scenarios well but typically lose nodal resolution and treat security as a post-hoc adjustment. Nodal optimal-power-flow tools (PowerWorld, DIgSILENT) handle security explicitly but are not built for stochastic multi-stage investment. Academic capacity-expansion frameworks (PyPSA, GenX, ReEDS) typically pick two of three (Rahmaniani et al., 2017). The result is that there is no widely-deployed, commercially-supported framework that simultaneously delivers nodal resolution, stochastic scenarios, and N-1 security at the scale infrastructure investors and system operators require.

Scope of this paper, and its place in the engine. The production response to that gap is the *Adaptive-Fidelity Nested Benders* (AFN-B) engine inside CENovaSage: a three-level nested decomposition in which an investment master is fed by per-period chronological recourses, with operational fidelity adapted as the solve converges in a way that keeps every accumulated cut a valid global minorant. AFN-B is the production core; its rigour and its convergence with the integer-recourse caveat are specified separately. *This paper specifies and verifies the single-level foundation* — the degenerate one-block case that AFN-B contains exactly when each period reduces to a single time block. We deliberately confine attention to that case so that two engineering questions can be settled in isolation, with full proofs and a reproducible numerical witness, before the nesting machinery is layered on: (i) how the cutting-plane master is stabilised, and (ii) how N-1 security is embedded without enumerating every contingency in the master.

The foundation rests on three pillars. *Benders decomposition* (Benders, 1962; Van Slyke and Wets, 1969) separates the two-stage problem into a master investment problem and per-scenario dispatch sub-problems linked by optimality cuts. *Stabilisation* replaces the unstable cutting-plane behaviour of naive Benders with a regularised iterate. Established stabilisers in the literature include in-out cut separation (Bonami et al., 2020) and proximal masters (Ruszczyński, 1986); in this paper we study two further *levers* on the same master — Magnanti–Wong Pareto cut selection (Magnanti and Wong, 1981) and a level-bundle quadratic master (Lemaréchal et al., 1995; Hiriart-Urruty and Lemaréchal, 1993; de Oliveira and Sagastizábal, 2014) — and compare them against the naive single-dual baseline. These are alternative cut/master configurations of the same algorithm, not a different algorithm. *N-1 separation* couples the dispatch sub-problem with a contingency oracle that screens post-contingency overloads by PTDF/LODF sensitivities and emits cuts only when a monitored corridor binds, rather than imposing all $|\mathcal{C}|$ contingencies as primal constraints in the master.

Our contribution is fourfold. First, we present the single-level foundation as a formal mathematical programme over a full DC power-flow network, and derive the LP-duality optimality cut together with the precise sense in which it is a valid global minorant under LP-relaxed operations (Sections 2–3). Second, we restate the convergence results at the rigour an expert

reader expects: finite termination of the mixed-integer master by a finitely-many-dual-bases argument, and the level-bundle rate stated qualitatively with its strong-convexity hypothesis explicitly flagged as non-binding for the piecewise-linear recourse here, the explicit constants relegated to Appendix B (Sections 3–4). Third, we describe the N-1 separation oracle, frame it honestly as security-constrained *economic* dispatch with VOLL-priced corrective actions, and bound the total number of cuts it can generate before finite termination (Section 5). Fourth, we demonstrate end-to-end correctness on a stylised GB 30-zone, 8-scenario instance engineered so that an N-1 boundary bottleneck genuinely binds, confirm that all six cut/master × oracle/enumeration configurations agree on the optimum to within the stated tolerance, and report the power-systems and optimality invariants (Section 6). We are explicit throughout about what the stylised instance can and cannot show: it is a correctness and mechanics demonstrator, not a performance benchmark, and the screening economy of the oracle requires a production-scale instance the companion work (in preparation) supplies.

Practitioner sidebar

What this means for capacity-expansion buyers. Existing commercial CEM tools force you to choose between resolution, security, and scenarios. A buyer who needs all three has had to either run them sequentially (zonal CEM with post-hoc security checking, with no convergence guarantee that the result is feasible) or hire bespoke modelling consultancies that re-implement the machinery from scratch. The CENovaSage engine integrates all three; this paper documents the foundation on which that engine is built and the discipline — valid bounds at every iterate, security priced rather than ignored — that makes its outputs auditable. The methodology is documented; the cost is no longer paid in rigour.

2 Problem formulation

2.1 Notation and decision variables

Let $\mathcal{N}$ index zones, $\mathcal{L}$ index transmission corridors, $\mathcal{K}$ index investment technologies, $\mathcal{S}$ index uncertainty scenarios with probabilities p_s summing to one, and $\mathcal{T} = \{1, \dots, T\}$ index operational time periods within a representative scenario. We denote the network incidence matrix $\boldsymbol{A} \in \{-1, 0, +1\}^{|\mathcal{L}|\times|\mathcal{N}|}$, with $A_{ln} = +1$ if line l is oriented out of zone n, -1 if into, 0 otherwise. The susceptance matrix $\boldsymbol{B}$ is diagonal with B_{ll} the per-unit susceptance of line l.

First-stage (investment) variables.

- $\boldsymbol{x}^{\text{gen}} \in \mathbb{R}_+^{|\mathcal{N}|\times|\mathcal{K}|}$: new generation/storage capacity by zone and technology (MW).
- $\boldsymbol{x}^{\text{line}} \in \{0, 1\}^{|\mathcal{L}|}$: binary line-reinforcement decisions (each line l, if reinforced, gains Δf_l MW of capacity).

Second-stage (operational) variables. For each scenario s and each operational period t:

- $g_{snkt} \in [0, (e^{\text{exist}}_{nk} + x^{\text{gen}}_{nk}) \cdot a_{snkt}]$: dispatch of technology k at zone n, where e^{exist}_{nk} is existing capacity and $a_{snkt} \in [0, 1]$ is availability.
- f_{slt}: power flow on line l, bounded by $\pm(f^{\max}_l + x^{\text{line}}_l \Delta f_l)$.
- θ_{snt}: voltage angle at zone n (one zone's angle is fixed as the slack reference).
- $r_{snt} \geq 0$: load shedding (involuntary curtailment of demand).

2.2 The full problem (deterministic equivalent)

The investment-and-dispatch problem is the deterministic equivalent of the two-stage stochastic programme:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \sum_{n,k} c^{\text{gen}}_{nk}\, x^{\text{gen}}_{nk} + \sum_{l} c^{\text{line}}_{l}\, x^{\text{line}}_{l} + \sum_{s} p_s\, Q_s(\boldsymbol{x}), \\ \text{s.t.} \quad & 0 \le x^{\text{gen}}_{nk} \le \overline{x}^{\text{gen}}_{nk}, \quad x^{\text{line}}_{l} \in \{0,1\}, \end{aligned} \tag{1}$$

where the operational cost-to-go function $Q_s(\boldsymbol{x})$ is the optimal value of the per-scenario dispatch LP of Section 2.3 (whose feasible set $\mathcal{Y}_s(\boldsymbol{x})$ internalises the base-case dispatch constraints); N-1 security enters through the security-constrained recourse $\widetilde{Q}_s$ of Section 2.4, which replaces Q_s throughout. We use $\boldsymbol{x} = (\boldsymbol{x}^{\text{gen}}, \boldsymbol{x}^{\text{line}})$ as shorthand for the joint first-stage decision.

2.3 Per-scenario dispatch sub-problem

For fixed $\boldsymbol{x}$, the per-scenario dispatch sub-problem in scenario s is the linear programme

$$\begin{aligned} Q_s(\boldsymbol{x}) := \min_{\boldsymbol{g},\boldsymbol{f},\boldsymbol{\theta},\boldsymbol{r}} \quad & \sum_{t}\sum_{n,k} \delta_t\, \nu_k g_{snkt} + \sum_{t}\sum_{n} \delta_t \nu^{\text{voll}} r_{snt} & & \\ \text{s.t.} \quad & \sum_{k} g_{snkt} - \sum_{l} A_{ln} f_{slt} + r_{snt} = D_{snt} & \forall t, n & \quad (\lambda_{snt}) \\ & f_{slt} = B_{ll} \sum_{n} A_{ln} \theta_{snt} & \forall t, l & \quad (\mu_{slt}) \\ & 0 \le g_{snkt} \le (e^{\text{exist}}_{nk} + x^{\text{gen}}_{nk}) a_{snkt} & \forall t, n, k & \quad (\sigma^{\text{gen}}_{snkt}) \\ & -(f^{\max}_{l} + \Delta f_l x^{\text{line}}_{l}) \le f_{slt} \le (f^{\max}_{l} + \Delta f_l x^{\text{line}}_{l}) & \forall t, l & \quad (\sigma^{\text{line}}_{slt}) \\ & 0 \le r_{snt} \le D_{snt} & \forall t, n & \end{aligned} \tag{2}$$

where δ_t is the duration weight of period t, ν_k is the variable cost of technology k (GBP/MWh), ν^{voll} is the value of lost load (GBP/MWh; a round 10 000 GBP/MWh in our calibration, within the range of published GB VoLL estimates), D_{snt} is demand, and the right-hand side notation (λ_{snt}), etc., denotes the dual variable associated with each constraint. The model enforces full DC power flow: bus voltage angles θ_{snt} with the flow–angle coupling $f_{slt} = B_{ll} \sum_n A_{ln} \theta_{snt}$ (one zone fixed as slack reference), exactly as in the production engine, not a transport (net-injection) relaxation. The framework extends to the AC second-order-cone relaxation (Low, 2014; Coffrin et al., 2016) with no change to the decomposition structure, though cut generation then rests on conic rather than LP duality (Section 7).

The dispatch sub-problem (2) is a finite linear programme; its optimal value $Q_s(\boldsymbol{x})$ is a convex piecewise linear function of $\boldsymbol{x}$, by parametric LP theory. This is the property we will use to construct Benders cuts.

2.4 N-1 security: cost-internalised corrective redispatch

We treat N-1 security as *security-constrained economic dispatch with VOLL-priced corrective actions*, not as a hard feasibility requirement. Let $\mathcal{C}$ be the contingency set; for each $c \in \mathcal{C}$ we write the post-contingency dispatch programme as $Q^c_s(\boldsymbol{x})$, identical in structure to (2) but with line c removed ($f^{\max}_c \to 0$) and with corrective re-dispatch and VOLL-priced non-served energy permitted. Base-case and post-contingency cuts share one epigraph variable per scenario, so the per-scenario operational value the master converges to is the worst case over the base and post-contingency dispatches,

$$\widetilde{Q}_s(\boldsymbol{x}) \;=\; \max\Big\{\, Q_s(\boldsymbol{x}),\; \max_{c \in \mathcal{C}} Q^c_s(\boldsymbol{x}) \Big\}. \tag{3}$$

Because every contingency dispatch carries an NSE term priced at VOLL, $\widetilde{Q}_s(\boldsymbol{x}) < \infty$ for all $\boldsymbol{x}$ (relatively complete recourse); a pointwise maximum of convex piecewise-linear functions is

convex piecewise-linear, so the cut machinery of Section 3 applies unchanged to $\widetilde{Q}_s$. Economically, an N-1 overload is therefore *priced* — a corridor trip that forces curtailment shows up as a VOLL-weighted cost the investment problem internalises — rather than excluded by a hard constraint.

The alternative, *hard* N-1 securing ($Q_s^c(\boldsymbol{x}) \le M_s$ for a feasibility threshold M_s, often zero permissible shed) is a one-line change to the separation logic of Section 5: replace the VOLL-priced post-contingency optimality cut by a Benders *feasibility* cut whenever the post-contingency shed exceeds M_s, exactly the feasibility-cut device that complete recourse otherwise makes unnecessary. We adopt the cost-internalised form throughout because it keeps recourse relatively complete (so no feasibility cuts are needed for convergence) and because pricing corrective actions is the economically meaningful object for an investment decision; the hard-feasibility variant is available where a regulatory standard mandates it. The conventional SCOPF formulation embeds all $|\mathcal{S}| \cdot |\mathcal{C}|$ post-contingency dispatch problems as additional constraint blocks in the master; the size of the resulting LP is the principal source of the intractability of full SCOPF-CEM.

3 Benders decomposition

3.1 Master and sub-problem split

Replace the cost-to-go terms $Q_s(\boldsymbol{x})$ by epigraph variables $\eta_s \ge Q_s(\boldsymbol{x})$ (we reserve θ for the bus voltage angles) and write the relaxed master:

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{\eta}} \quad & \sum_{n,k} c^{\text{gen}}_{nk} x^{\text{gen}}_{nk} + \sum_{l} c^{\text{line}}_{l} x^{\text{line}}_{l} + \sum_{s} p_s \eta_s \\ \text{s.t.} \quad & \eta_s \ge \alpha_s^{(j)} + \langle \boldsymbol{\beta}_s^{\text{gen},(j)}, \boldsymbol{x}^{\text{gen}} \rangle + \langle \boldsymbol{\beta}_s^{\text{line},(j)}, \boldsymbol{x}^{\text{line}} \rangle \quad \forall (s,j) \in \mathcal{J}_k \\ & 0 \le \boldsymbol{x}^{\text{gen}} \le \overline{\boldsymbol{x}}^{\text{gen}},\ \boldsymbol{x}^{\text{line}} \in \{0,1\}^{|\mathcal{L}|}, \end{aligned} \tag{4}$$

where $\mathcal{J}_k$ is the set of cuts accumulated through outer iteration k. Each cut is generated by solving the dispatch sub-problem (2) at a trial point $\hat{\boldsymbol{x}}$ and using LP duality to extract the affine lower bound:

$$\begin{aligned} \alpha_s &= Q_s(\hat{\boldsymbol{x}}) - \langle \boldsymbol{\beta}_s^{\text{gen}}, \hat{\boldsymbol{x}}^{\text{gen}} \rangle - \langle \boldsymbol{\beta}_s^{\text{line}}, \hat{\boldsymbol{x}}^{\text{line}} \rangle, \\ \beta^{\text{gen}}_{s,nk} &= -\sum_{t} \sigma^{\text{gen}}_{snkt}\, a_{snkt}, \\ \beta^{\text{line}}_{s,l} &= -\sum_{t} \sigma^{\text{line}}_{slt}\, \Delta f_l, \end{aligned} \tag{5}$$

where the duals σ^{gen}, σ^{line} are those of the capacity-bound constraints in (2), whose objective already carries the duration weight δ_t — the duals therefore carry it too, and no additional δ_t factor may appear in the coefficients (double-counting it would steepen the cut by δ_t and destroy the minorant property for non-uniform period weights). The cut coefficients aggregate these duals by their multiplicative dependence on $\boldsymbol{x}$. Existence and uniqueness of these duals follow from the boundedness and feasibility of the dispatch LP.

3.2 Algorithm: classical (naive) Benders

Algorithm 1 Naive Benders for stochastic CEM

Require: Problem data, tolerance $\varepsilon > 0$, max iter K
Ensure: Optimal $\boldsymbol{x}^*$ and value z^*
Initialise $\mathcal{J}_0 \leftarrow \emptyset$, $\hat{\boldsymbol{x}}_0 \leftarrow \mathbf{0}$
for $k = 0, 1, \ldots, K-1$ **do**
  Solve master (4) with cuts $\mathcal{J}_k$, get $(\boldsymbol{x}_k, \boldsymbol{\eta}_k)$
  Set $\mathrm{LB}_k = \langle \boldsymbol{c}, \boldsymbol{x}_k \rangle + \sum_s p_s \eta_{s,k}$, the *unregularised* cut-model optimum ▷ bound-trap discipline, Rem. 3.2
  for $s \in \mathcal{S}$ **do**
    Solve dispatch (2) at $\boldsymbol{x}_k$, get $Q_s(\boldsymbol{x}_k)$ and duals
    Append cut $(\alpha_s, \boldsymbol{\beta}_s^{\mathrm{gen}}, \boldsymbol{\beta}_s^{\mathrm{line}})$ to $\mathcal{J}_{k+1}$
  end for
  Compute $\mathrm{UB}_k = \langle \boldsymbol{c}, \boldsymbol{x}_k \rangle + \sum_s p_s Q_s(\boldsymbol{x}_k)$ at *exact* fidelity
  if $(\mathrm{UB}_k - \mathrm{LB}_k)/|\mathrm{UB}_k| < \varepsilon$ **then break**
  end if
end for
return $\boldsymbol{x}_k, \mathrm{UB}_k$

3.3 Validity of the cut and the bound-trap discipline

We first record the property that underwrites every bound in this paper. With LP-relaxed operations, Q_s (and $\widetilde{Q}_s$ of (3)) is convex piecewise-linear in $\boldsymbol{x}$, and the cut (5) is the supporting hyperplane at $\hat{\boldsymbol{x}}$; it is therefore a valid global minorant of Q_s everywhere and tight at $\hat{\boldsymbol{x}}$. The same property underpins the cut analysis of Cole et al. (2026, §4), and it is what makes the master lower bound valid.

Remark 3.1 (Integer recourse caveat). The validity of (5) as a global minorant relies on the operational sub-problem being a linear programme. If the *exact* operational fidelity contained unit-commitment binaries, Q_s would be non-convex in $\boldsymbol{x}$ and the LP-duality cut would no longer be a valid global underestimator. In this paper, and in the engine's investment loop, operations are LP-relaxed for cut generation precisely to preserve validity; unit-commitment integrality, where required, is handled either by a final operational-validation pass with the integrality gap reported, or by Lagrangian (SDDiP) cuts that are valid for integer recourse (Zou et al., 2019). One must not silently apply an LP-duality cut to a binary operational model and call the resulting bound valid.

Remark 3.2 (Bound-trap discipline). The lower bound must always be read from the *unregularised* cut model (4), $\mathrm{LB}_k = \langle \boldsymbol{c}, \boldsymbol{x}_k \rangle + \sum_s p_s \eta_{s,k}$, never from a regularised objective. The proximal term and the level constraint of the regularised master (6) both perturb the objective and do not lower-bound the optimal value; reporting the regularised objective as the lower bound is a silent correctness bug that inflates the apparent gap and can stop the algorithm early. Algorithm 1 computes LB_k from the cut-model optimum for this reason, and the level-bundle variant of Section 4 re-reads its lower bound from the same unregularised model. This discipline is the single most error-prone point in a regularised or adaptive implementation.

3.4 Convergence of naive Benders

Theorem 3.3 (Finite convergence, integer master)**.** *Suppose the first-stage set is bounded with integer line decisions* $\boldsymbol{x}^{line} \in \{0,1\}^{|\mathcal{L}|}$*, the dispatch sub-problem* (2) *has relatively complete recourse, no cut is deleted from the store, and the master is re-solved to global optimality at each major iteration. Then Algorithm 1 terminates finitely at the optimal value of* (1)*.*

Proof. This is the classical finite-basis argument of Van Slyke and Wets (1969), and it is the argument the *mixed*-integer first stage requires: because $\boldsymbol{x}^{\text{gen}}$ is continuous, the same integer point $\boldsymbol{x}^{\text{line}}$ can be revisited with a different $\boldsymbol{x}^{\text{gen}}$ at which the earlier cuts are slack, so a per-integer-point count does not bound the major iterations. What is finite is the cut alphabet. Each dispatch LP (2) has finitely many dual basic solutions, and every optimality cut (5) is generated from one of them, so only finitely many distinct cuts exist. At a non-terminal major iteration the master value at $\hat{\boldsymbol{x}}_k$ satisfies $\sum_s p_s \eta_{s,k} < \sum_s p_s Q_s(\hat{\boldsymbol{x}}_k)$ (otherwise $\text{LB}_k \geq \text{UB}_k$ and the algorithm stops), so for at least one scenario the freshly generated cut is violated by $(\hat{\boldsymbol{x}}_k, \boldsymbol{\eta}_k)$ and is therefore not already in the store: each non-terminal iteration adds at least one previously-unseen cut. Since cuts are never deleted, the algorithm terminates after finitely many major iterations, and at termination LB = UB certifies optimality (cf. Sherali and Fraticelli, 2002; Rahmaniani et al., 2017). Integrality of the line decisions is *not* what drives this count; its role is to make the set $\mathcal{V}_{\text{int}}$ of Theorem 5.1 finite. □

Remark 3.4 (No geometric rate in the continuous relaxation). When the line decisions are relaxed to $\boldsymbol{x}^{\text{line}} \in [0,1]^{|\mathcal{L}|}$, the master is a continuous cutting-plane method on the convex piecewise-linear function $\sum_s p_s Q_s$. We do *not* claim a geometric rate for it. A geometric contraction would require $\sum_s p_s Q_s$ to be strongly convex, and *strong convexity fails identically for a piecewise-linear function*: the recourse here is polyhedral, so no $\mu > 0$ exists. The relevant guarantee is therefore the worst-case behaviour of Kelley's method (Kelley, 1960; Rahmaniani et al., 2017) — finite termination at a vertex-characterising cut set, but with the well-documented *tailing-off*: rapid initial gap reduction followed by slow approach. This tailing-off, not a provable asymptotic-rate gap, is what the stabilisation levers of Section 4 are designed to mitigate.

4 Stabilisation: Magnanti–Wong and level-bundle

4.1 Magnanti–Wong cut selection

When the dispatch sub-problem is degenerate (multiple dual optima), naive Benders selects an arbitrary dual and produces an arbitrary cut. The cut is valid but may not be *Pareto-optimal* in the sense of dominating other cuts that could have been generated from the same primal solution. Magnanti and Wong (1981) show that selecting the dual $\boldsymbol{\sigma}^*$ that maximises $\langle \boldsymbol{\sigma}, \boldsymbol{x}^0 - \hat{\boldsymbol{x}} \rangle$ for a fixed core point $\boldsymbol{x}^0$ in the master feasible set produces a Pareto-optimal cut, and that this is implementable as a second LP solve over the dual polyhedron.

The Magnanti–Wong cut is generated by:

1. Solve dispatch (2) at $\hat{\boldsymbol{x}}$, recording optimal value $Q_s(\hat{\boldsymbol{x}})$.
2. Among the dual optima, select $\boldsymbol{\sigma}^*$ maximising $\langle \boldsymbol{\sigma}, \boldsymbol{x}^0 - \hat{\boldsymbol{x}} \rangle$ where $\boldsymbol{x}^0$ is a relative interior point of the master feasible region. This is the second LP.
3. Form the cut from (5) using $\boldsymbol{\sigma}^*$.

Magnanti–Wong cut selection requires an auxiliary dual LP per cut, evaluated at the interior core point, so it trades a heavier per-iteration solve for tighter cuts; whether that trade pays depends on how degenerate the sub-problems are and on how many major iterations the tighter cuts save (Rahmaniani et al., 2017; Muñoz et al., 2016). We make no quantitative speed-up claim here — on the stylised instance below the auxiliary solve roughly doubles the naive wall time, and the instance converges in too few iterations for the tighter cuts to amortise (Section 6).

4.2 Level-bundle stabilisation

Level-bundle methods (Lemaréchal et al., 1995; Hiriart-Urruty and Lemaréchal, 1993; de Oliveira and Sagastizábal, 2014) address the master-instability issue at its root: rather than letting the

master jump freely to the optimum of the cut polyhedron at each iteration, the master is required to stay within a level set of the past iterates and a stability centre. The level-bundle master studied here replaces (4) with:

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{\eta}} \quad & \tfrac{1}{2}\|(\boldsymbol{x}^{\mathrm{gen}}, \boldsymbol{x}^{\mathrm{line}}) - \boldsymbol{x}^{\mathrm{cen}}\|_2^2 \\ \text{s.t.} \quad & \eta_s \geq \alpha_s^{(j)} + \langle \boldsymbol{\beta}_s^{(j)}, \boldsymbol{x}\rangle, \quad (s,j) \in \mathcal{J}_k \\ & \langle \boldsymbol{c}, \boldsymbol{x}\rangle + \sum_s p_s \eta_s \leq L_k, \\ & 0 \leq \boldsymbol{x}^{\mathrm{gen}} \leq \overline{\boldsymbol{x}}^{\mathrm{gen}}, \ \boldsymbol{x}^{\mathrm{line}} \in [0,1]^{|\mathcal{L}|}. \end{aligned} \tag{6}$$

The level $L_k = \mathrm{LB}_k + \kappa(\mathrm{UB}_k - \mathrm{LB}_k)$ for $\kappa \in (0,1)$ defines a slice of the cut polyhedron; we use $\kappa = 0.6$ throughout the case study (Remark 4.3), and crucially LB_k is the unregularised cut-model bound of Remark 3.2, never the regularised objective. The stability centre $\boldsymbol{x}^{\mathrm{cen}}$ is updated to the incumbent best whenever a sufficient-descent test is met (a serious step), otherwise the centre is retained and the freshly generated cuts are kept (a null step). The objective is the squared distance to the centre over the investment variables $(\boldsymbol{x}^{\mathrm{gen}}, \boldsymbol{x}^{\mathrm{line}})$ — the projection step of the level method; a multiplicative weight on a pure quadratic objective would not change the minimiser, so none appears, and the epigraph variables are deliberately excluded from the distance (penalising $\boldsymbol{\eta}$, whose scale is the operating cost, would swamp the investment term by orders of magnitude). This is a convex quadratic programme; we solve it with CLARABEL in the case study (OSQP as fallback). It is one configuration of the same master, not a separate algorithm.

Theorem 4.1 (Level-bundle convergence rate, qualitative)**.** *Let $f := \langle \boldsymbol{c}, \cdot\rangle + \sum_s p_s Q_s$ be convex and L-Lipschitz on the master feasible region, of diameter D, and let $\kappa \in (0,1)$ be the level fraction in (6). Then the level-bundle iterates converge to the optimum, and the optimality gap decays at the rate $O(1/\sqrt{k})$,*

$$\mathrm{UB}_k - z^* \;=\; O\!\left(\frac{DL}{\sqrt{k}}\right), \qquad k \geq 1, \tag{7}$$

which is the uniformly optimal worst-case rate for a black-box first-order method on a non-smooth Lipschitz convex problem.

Proof. This is the level-method rate of Lemaréchal et al. (1995, Thm. 3.5); see also the textbook treatment of Bonnans et al. (2006, Thm. 9.2.2) and the on-demand-accuracy analysis of de Oliveira and Sagastizábal (2014). Optimality of the $O(1/\sqrt{k})$ rate among black-box first-order methods on non-smooth Lipschitz convex problems is the lower bound of Lan (2015), who shows that bundle-level methods attain it. The explicit value of the rate constant, with its dimensional derivation, is given in Appendix B. □

Remark 4.2 (Strong convexity fails here; the geometric bound is a non-binding hypothesis). A sharper, *geometric* level-bundle rate is available under the additional hypothesis that f is μ-strongly convex for some $\mu > 0$ (de Oliveira and Sagastizábal, 2014, Sec. 3.4). That hypothesis *does not hold for the problem of this paper.* The recourse Q_s (and $\widetilde{Q}_s$) is convex *piecewise-linear* under LP-relaxed operations, and a piecewise-linear function is not strongly convex anywhere; the first-stage cost is linear. Hence no $\mu > 0$ exists and the geometric bound is vacuous here. We state the rate qualitatively in (7) for exactly this reason: the $O(1/\sqrt{k})$ bound is the binding guarantee, and any geometric improvement would require a strongly-convex regulariser deliberately added to the objective (which the proximal term does *per-iteration*, but which does not change the value-function geometry that governs the asymptotic rate). We flag this because conflating the Lipschitz-convex and strongly-convex regimes is a common over-claim, and an expert reader will rightly object to a geometric rate asserted over piecewise-linear recourse.

Remark 4.3 (Choosing κ). The rate constant in (7) (Appendix B) carries a factor $1/\sqrt{2\kappa(1-\kappa)}$, minimised at $\kappa = 1/2$. Asymmetric choices $\kappa \in (0.5, 0.7)$ are common in practice because they stabilise behaviour when the oracle returns inexact cuts (de Oliveira and Sagastizábal, 2014); we use $\kappa = 0.6$ in the case study, for which $1/\sqrt{2 \cdot 0.6 \cdot 0.4} \approx 1.44$, about 2% above the $\kappa = 0.5$ value of $\sqrt{2}$.

Remark 4.4 (What stabilisation buys, stated honestly). On a piecewise-linear convex f, naive Benders carries no per-iteration progress guarantee — the gap can stagnate for many consecutive iterations before dropping sharply (the *tailing-off* phenomenon, Rahmaniani et al. 2017). Theorem 4.1 gives the level-bundle master a guaranteed $O(1/\sqrt{k})$ gap decay from iteration 1, which is what removes tailing-off in practice. This is an *asymptotic worst-case* statement: on a small, easy instance — such as the stylised case study below, which all configurations solve in three major iterations — it predicts no advantage, and indeed none is observed. The value of stabilisation is realised on large, ill-conditioned instances with many major iterations; demonstrating it quantitatively is the role of the production-scale companion study, not of this foundation paper. We deliberately make no speed-up claim at the stylised scale.

4.3 Combining Magnanti–Wong with level-bundle

Magnanti–Wong cut selection and level-bundle stabilisation are orthogonal — the former is a property of the cut, the latter of the master — and compose: one can generate Pareto-optimal cuts and feed them to the level-bundle QP. The combined configuration inherits the rate of Theorem 4.1, with the cut-selection rule reducing the number of cuts the master must process per iteration at the cost of the auxiliary dual solve. We report both stabilisers separately rather than only the combination, so that the contribution of each lever is visible where cut quality or master conditioning limits convergence.

5 The N-1 separation oracle

5.1 Why SCOPF-master is intractable

The full SCOPF master embeds the post-contingency dispatch problems as constraint blocks in the master (or, in the decomposed form, evaluates every (s, c) pair at every trial point), requiring per-scenario per-contingency dispatch variables. The number of additional variables and constraints scales as $|\mathcal{S}| \cdot |\mathcal{C}| \cdot T \cdot |\mathcal{N}|$. For a production-scale instance — tens of scenarios, hundreds of contingencies, a full chronological year at nodal resolution — this product runs into the billions of variables; even with state-of-the-art LP solvers and decomposition the memory footprint alone is prohibitive.

The observation underlying the oracle is that, at a production scale where $|\mathcal{C}|$ is large, only a small subset of contingencies bind at the optimum; the rest are slack and need not trigger a full post-contingency solve. The N-1 separation oracle exploits this by screening with PTDF/LODF sensitivities and solving the full post-contingency dispatch only for screened candidates. We stress that this screening *economy* is a property of large, sparsely-binding contingency sets: the stylised instance of Section 6 is constructed so that *all* of its monitored corridors bind, precisely so that the oracle's correctness can be checked against full enumeration on every contingency; it therefore cannot, and is not used to, demonstrate the screening economy. That demonstration belongs to the production-scale companion work (in preparation).

5.2 The oracle

Algorithm 2 N-1 separation oracle

Require: Trial investment $\hat{\boldsymbol{x}}$, scenario s, candidate set $\mathcal{C}$
Ensure: Either "feasible" or a contingency-aware cut
1: Solve base dispatch (2) at $\hat{\boldsymbol{x}}$, get base flows $\boldsymbol{f}^*$
2: *(Pre-screen)* For each $c \in \mathcal{C}$, estimate post-contingency flows by the line-outage distribution factors (LODF) Ψ: $\hat{f}_l^c = f_l^* + \Psi_{lc}\, f_c^*$, $l \neq c$
3: Identify $\mathcal{C}^{\text{cand}} \subseteq \mathcal{C}$: contingencies for which $|\hat{f}_l^c|$ exceeds γ times line capacity for some l ▷ screen margin γ; $\gamma = 0.9$ in the case study
4: **for** $c \in \mathcal{C}^{\text{cand}}$, worst-first, at most $n_{\max}$ per call **do** ▷ $n_{\max} = 8$ in the case study
5: Solve full post-contingency dispatch (2) with line c removed
6: **if** post-contingency cost exceeds the base cost beyond tolerance τ **then**
7: Generate a VOLL-priced post-contingency optimality cut from its duals
8: **end if**
9: **end for**
10: Emit all generated cuts (or "no violation" if $\mathcal{C}^{\text{cand}} = \emptyset$ or none bind)
11: *(Fallback)* If the master gap ever turns negative — a screen miss — the driver re-runs one iteration with full enumeration of $\mathcal{C}$, restoring a valid upper bound

The pre-screen uses the line-outage distribution factors Ψ (computed from the network's PTDF matrix, themselves derived from the susceptances and incidence from first principles) to identify candidate contingencies in $O(|\mathcal{L}|)$ time per contingency, avoiding an LP solve for every $c \in \mathcal{C}$. Only $\mathcal{C}^{\text{cand}}$ triggers the full post-contingency LP. The screen margin γ, the per-call solve cap $n_{\max}$ and the enumeration fallback are settings of this case study, not engine settings; $\gamma < 1$ catches contingencies that bind in *cost* (forcing expensive redispatch) before they bind thermally, and the fallback guarantees correctness if the screened-and-capped economy ever misses a binding contingency. The economy of the screen is the ratio $|\mathcal{C}^{\text{cand}}|/|\mathcal{C}|$; note that at production scale, where the cap can bind, the number of post-contingency LPs actually solved is governed by $n_{\max}$, so the screen ratio is not identical to the ratio of LP solves avoided. We make no numerical speed-up claim here, because on the stylised instance the screen ratio is exactly 1.0 (every monitored corridor binds, by construction); a small ratio, and the corresponding per-iteration saving over full enumeration, requires a production-scale contingency set and is reported in the companion work (in preparation).

5.3 Convergence with the oracle

Theorem 5.1 (Finite termination with the N-1 oracle)**.** *Under the assumptions of Theorem 3.3, with N-1 handled in the cost-internalised form of Section 2.4, and assuming additionally that every $c \in \mathcal{C}$ leaves the network connected (so the LODF column $\Psi_{\cdot c}$ is well defined) and that the cut tolerance is $\tau = 0$, Algorithm 1 augmented with the oracle (Algorithm 2) generates at most*

$$K\,|\mathcal{S}|\,(1 + |\mathcal{C}|)$$

cuts over its K major iterations — at most one base optimality cut per (scenario, iteration) pair plus at most one post-contingency optimality cut per (scenario, contingency, iteration) triple — with K finite by Theorem 3.3; when the first stage is purely integer, $K \leq |\mathcal{V}_{int}|$ and the bound reads $|\mathcal{V}_{int}|\,|\mathcal{S}|\,(1 + |\mathcal{C}|)$. The algorithm therefore terminates finitely, returning the same optimal value as enforcing every contingency by enumeration (for $\tau > 0$, the same value up to $O(\tau)$).

Proof. Per major iteration, each scenario contributes at most one base optimality cut, and each (scenario, contingency) pair at most one post-contingency cut, giving at most $|\mathcal{S}|(1 + |\mathcal{C}|)$ cuts

per iteration and $K|\mathcal{S}|(1+|\mathcal{C}|)$ in total; K is finite by Theorem 3.3 (whose no-deletion hypothesis we inherit — a pruned store preserves the validity of the lower bound but not its monotonicity, so pruning voids the count, not the correctness[1]), and when the first stage is purely integer the argument of one major iteration per integer point bounds K by $|\mathcal{V}_{\text{int}}|$. The oracle returns the enumeration optimum because a contingency that the pre-screen passes is non-binding at $\hat{\boldsymbol{x}}$ at the LODF estimate — exact for DC flow at fixed base injections when the outage leaves the network connected, which the connectedness hypothesis guarantees (islanding outages are radial elements excluded from $\mathcal{C}$ by construction) — so omitting its cut cannot remove the optimum, while every binding contingency is cut exactly as under enumeration; with $\tau > 0$ a contingency whose cost exceedance is below τ may go uncut, which perturbs the optimum by at most $O(\tau)$. □

Remark 5.2 (The bound is loose, and we do not over-read it). The bound $K|\mathcal{S}|(1 + |\mathcal{C}|)$ is a worst case; in practice far fewer cuts are generated because few contingencies bind. On the stylised instance the count is exact — the naive run stores $96 = 3 \times 8 \times 4$ cuts over its three iterations — and the oracle and the full enumeration agree on the optimum to machine precision, generating the same 72 post-contingency cuts (Section 6); they are validated as equivalent there. The per-iteration wall-time saving the oracle confers over full enumeration at production scale is a separate, quantitative claim that this instance does not support and that we therefore defer.

Practitioner sidebar

Why this matters operationally. A capacity expansion plan that ignores N-1 looks fine in the modelled outcome but creates real-world reliability risk: when a corridor trips, the system's ability to redispatch is curtailed, and an unsecured CEM never sees the resulting cost. Embedding N-1 in the planning loop — here by *pricing* corrective redispatch and contingency-driven curtailment at the value of lost load, so the investment problem trades reinforcement against worst-case post-contingency cost — internalises that risk into the capital decision. Investment plans signed off on un-secured CEM can result in commissioned-but-unusable transmission, stranded assets, and emergency redispatch costs. Where a regulatory standard requires *hard* security rather than priced security, the feasibility-cut variant of Section 2.4 enforces it. Pricing or enforcing N-1 in the planning loop is not optional for investment-grade work; it is what investment-grade means.

6 Stylised GB 30-zone case study

6.1 Setup

We illustrate the foundation on a synthetic 30-zone network with a GB-like north–south split (generation-rich north, demand-heavy, build-constrained south). The network has 47 transmission corridors — the 49 edges of a meshed 6×5 grid less the two boundary crossings dropped to leave a three-corridor cut — deliberately engineered with a single narrow north–south bottleneck: the two halves are joined only by a "boundary-B6" cut of *three* tightly-rated parallel corridors. Investment options at each zone span four candidate technologies (wind, solar, gas-CCGT, and a battery-like flexible resource modelled without state-of-charge dynamics at this scale). Eight scenarios of eight representative hours each capture weather and load variation. The monitored contingency set is exactly these three boundary corridors, $|\mathcal{C}| = 3$: tripping any one drops the boundary to two circuits and overloads the survivors, so the N-1 constraint is genuinely binding at the optimum. The DC-OPF subproblem is solved from first principles — bus angles, susceptances, slack reference, and PTDF/LODF sensitivities all derived from the network data — with dispatch LP solves taking a median of ≈9 ms (95th percentile 10–14 ms), measured in the

[1]The implementation caps the store at 60 cuts per scenario; the cap is not reached on the naive and Magnanti–Wong runs of Section 6 and is reached only in the final iterations of the level-bundle runs.

same runs that produce Table 1 and consistent with its wall times. The convergence tolerance is $\varepsilon = 10^{-5}$ (relative).

The stylisation is deliberate and its limits are stated up front. The instance exists to demonstrate *correctness and mechanics* end-to-end in a reproducible script: that the real DC power flow, the genuine Magnanti–Wong auxiliary solve, the genuine level-bundle QP with the bound-trap discipline, and the firing PTDF/LODF N-1 oracle all behave as specified and agree on a single optimum — and each of those "genuine"s is *certified from the run's own trace*, not asserted: the level-QP execution and its per-iteration movement of the trial point, the Pareto-versus-fallback outcome of every auxiliary solve, and a post-hoc validity audit of every stored cut are recorded and checked mechanically (Section 6.3). It is *not* a performance benchmark, and two consequences follow that we do not paper over. First, the first-stage line builds are LP-relaxed here ($\boldsymbol{x}^{\text{line}} \in [0,1]$; production keeps them binary), so the integer-master iteration count of Theorem 3.3 is not exercised. Second, the instance is small enough that the naive and Magnanti–Wong configurations converge in three major iterations — the level-bundle master takes seventeen, its deliberately conservative level steps being the cost of stabilisation machinery on an instance with no tailing-off to remove — so it cannot surface any stabilisation speed-up (Remark 4.4), and its monitored contingencies all bind ($|\mathcal{C}^{\text{cand}}|/|\mathcal{C}| = 1.0$), so it cannot surface the oracle's screening economy (Remark 5.2). Those quantitative properties belong to a production-scale instance with binary builds and a far larger, sparsely-binding contingency set, reported in the nested-engine companion work (in preparation; Section 6.4).

6.2 Correctness check across cut/master and security configurations

Table 1 reports the converged optimum, major-iteration count, wall-clock time, and the N-1 separation statistics for all six configurations: the three cut/master choices (naive single-dual cuts; Magnanti–Wong Pareto cuts; level-bundle QP) crossed with the two security implementations (full enumeration, evaluating every contingency at every trial point; and the PTDF/LODF oracle). All six agree on the optimum of £74.67M (£74,665,367.2): the naive and Magnanti–Wong configurations to machine precision, the level-bundle configurations to a relative spread of 8.1×10^{-6}, inside their 10^{-5} termination tolerance. The agreement is a genuine cross-validation, not a re-run of one computation. The level-bundle trajectory is distinct from the naive one at every iterate — certified from the run's trace: the level-QP executed on all sixteen eligible iterations, moved the trial point every time, and held its level constraint active — while the Magnanti–Wong configurations reproduce the naive trajectory for an instructive reason: at the trial points this instance visits, the dispatch duals are essentially unique, so the Pareto-selected dual coincides with the naive dual to within 5×10^{-7} in cut coefficients. The auxiliary machinery is fully exercised (24 of 24 auxiliary solves returned the genuine Pareto cut; the validity guard never fired); its selection differs from the naive cut only where dual degeneracy exists to exploit. The naive and Magnanti–Wong masters converge in three major iterations; the level-bundle master takes seventeen.

The wall times carry the honest message of this paper. The Magnanti–Wong auxiliary dual LP at the core point is real work: it roughly doubles the naive wall time (≈ 3.2 s versus ≈ 1.6 s) without changing the iteration count, and at three iterations there are too few major steps for its tighter cuts to amortise. The level-bundle configuration is slower still ($\approx$ 11–12 s): its conservative level steps take seventeen major iterations where the naive master takes three. None of this is a stabilisation *speed-up*; on an instance this small there is no tailing-off to remove, and we claim none (Remark 4.4). The oracle and full-enumeration columns differ negligibly in wall time here because $|\mathcal{C}^{\text{cand}}|/|\mathcal{C}| = 1.0$: every contingency is screened in and solved, so the oracle does the same post-contingency solves as the enumeration (72 in the naive runs). That equality is exactly why the comparison is a clean correctness check rather than a performance result — and the enumeration baseline is the same decomposition with the screen disabled, not the constraint-embedded SCOPF master of Section 5, which this instance is too small to require.

Table 1: Configuration comparison on the stylised GB 30-zone, 8-scenario, $|\mathcal{C}| = 3$ instance. All six configurations agree on the optimum, £74.67M, to a relative spread of 8.1×10^{-6} (naive and Magnanti–Wong to machine precision). Wall times reflect that the Magnanti–Wong auxiliary solve and the level-bundle QP are genuine work that does not pay off at this scale, and that the oracle and full enumeration do identical work because every monitored corridor binds ($|\mathcal{C}^{\text{cand}}|/|\mathcal{C}| = 1.0$). This is a correctness/mechanics demonstrator; it is not, and is not presented as, a performance benchmark. Source: `results/variant_comparison.json`.

Configuration	Iters	Wall (s)	Optimum (£M)	N-1 cuts
Naive + full enumeration	3	1.65	74.67	72
Naive + N-1 oracle	3	1.58	74.67	72
Magnanti–Wong + full enumeration	3	3.13	74.67	72
Magnanti–Wong + N-1 oracle	3	3.23	74.67	72
Level-bundle + full enumeration	17	10.99	74.67	408
Level-bundle + N-1 oracle	17	11.73	74.67	408

Figure 1 shows the bound trajectories closing to the common optimum; Figure 2 shows the wall-time and iteration breakdown that makes the stabiliser overheads visible.

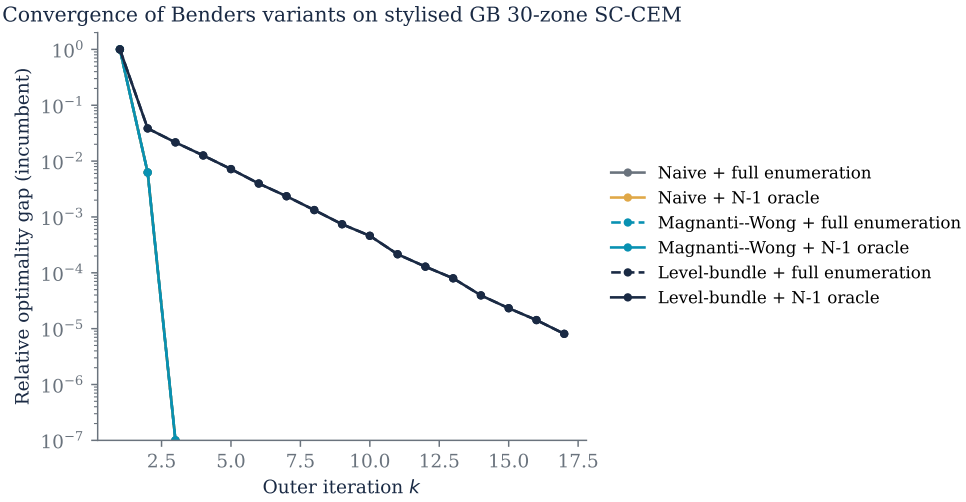

Figure 1: Convergence of the incumbent relative gap across all six configurations on the stylised instance. The naive and Magnanti–Wong curves coincide (the dispatch duals are essentially unique at the visited iterates, so Pareto selection returns the naive cut); the level-bundle curves are distinct, closing over seventeen conservative level steps. All configurations close on the common optimum £74.67M; the lower bound is read from the unregularised cut model in every case (the bound-trap discipline of Remark 3.2).

6.3 N-1 mechanics and invariant checks

The N-1 oracle genuinely fires. Across the naive run, 72 contingency cuts are generated and 72 post-contingency DC-OPF problems are solved (408 across the seventeen level-bundle iterations); the worst contingency (the trip of a boundary corridor) produces an LODF overload factor of 1.64 on a surviving circuit and a scenario-0 post-contingency cost of £98.7M, priced through the VOLL term. The oracle and the full enumeration agree on the optimum to machine precision, validating the screening logic as an exact substitute on this instance. Table 2 scopes every price and shed figure quoted in this section to the population it is computed over. The optimum accepts a priced, probability-weighted *base-case* non-served energy of 1,756 MWh (by scenario,

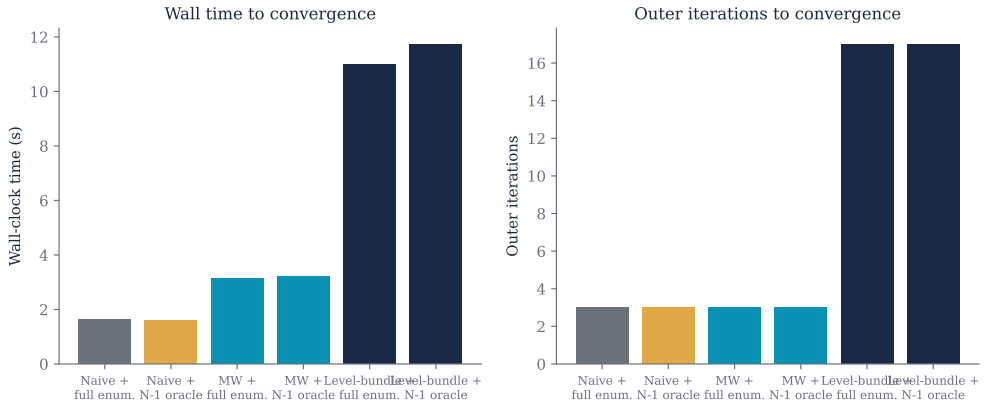

Figure 2: Wall-clock time and outer iterations by configuration. Magnanti–Wong roughly doubles the naive wall time (a genuine auxiliary dual solve per cut, at the core point); the level-bundle configurations are slowest because their conservative level steps take seventeen iterations against the naive three. The oracle and full-enumeration pairs are near-identical because all three monitored corridors bind. No stabilisation speed-up is claimed at this scale.

500–3,588 MWh): at the stylised eight-representative-hour operational slice, annualised capacity cost exceeds the VOLL cost of shedding in the rarest peak hours, so the optimum rationally sheds a little rather than overbuild. Against that, the worst single contingency forces 9,631 MWh of shed in scenario 0 (by scenario, 3,614–9,631 MWh): the N-1 event, not the base case, dominates the priced security cost, which is precisely the bottleneck the instance was engineered to exhibit.

The power-systems and optimality invariants all pass to numerical tolerance:

- **Nodal power balance:** maximum nodal residual 2.65×10^{-10} MW across all eight scenarios' base-case dispatches, every hour (tolerance 10^{-3} MW); the post-contingency dispatches are re-solved and checked by the scope-table audit of Table 2.
- **DC flow consistency:** $f_l = B_l(\theta_{\text{from}} - \theta_{\text{to}})$ holds by construction on every in-service line.
- **Thermal limits:** maximum loading $|f|/f^{\max} = 1.0000$ across all scenarios' base cases (≤ 1; the boundary corridors bind at the optimum, as designed).
- **Prices:** LMPs equal the duals of nodal balance. Scenario-0 base case: mean +2,414, minimum −7,625, maximum +10,000 £/MWh; across all scenarios and contingencies prices span −8,610 to +10,000 £/MWh (Table 2). The +10,000 ceiling is the VoLL, attained exactly at nodes that shed; the deeply negative LMPs arise under congestion — an increment of demand at such a node relieves a VOLL-scale binding constraint elsewhere — and are correct, not a sign error.
- **Cut validity:** every stored cut across all six configurations (1,344 cuts) is verified *post hoc* to be a valid global minorant of its recourse (base-case or post-contingency) on 50 uniformly-drawn first-stage points — a shared table of 1,600 recourse LP solves — and tight at its generating point against an independent re-solve: maximum minorant violation 0.0, maximum tightness error 3.0×10^{-16} relative.

6.4 What the small-scale case study does and does not show

To be unambiguous about scope: the stylised instance establishes that the mechanics are correct — real DC power flow, genuine Pareto and level-bundle stabilisation with valid bounds, a firing N-1 oracle, all six configurations agreeing on one optimum with every invariant satisfied. It establishes *nothing* about performance, by design. It cannot show a stabilisation speed-up (the

Table 2: Price and non-served-energy figures with their scopes, at the converged plan. Every prose number in this section corresponds to a row. LMPs in £/MWh; NSE in MWh. Source: `results/price_nse_scope.json`.

Quantity and scope	**Min**	**Mean**	**Max**
LMP, scenario-0 base case	−7,625	+2,414	+10,000
LMP, all scenarios, base case	−8,610	—	+10,000
LMP, all scenarios, post-contingency	−7,723	—	+10,000
NSE, base case, by scenario	500	—	3,588
NSE, base case, probability-weighted total			1,756
NSE, worst single contingency, by scenario	3,614	—	9,631
NSE, worst single contingency, scenario 0			9,631

naive master takes three iterations and the level-bundle takes seventeen; both stabilisers are in fact slower in wall time), and it cannot show the oracle's screening economy (its monitored contingencies all bind, so $|\mathcal{C}^{\text{cand}}|/|\mathcal{C}| = 1.0$).

Those quantitative questions require the production-scale CENovaSage instance — binary line decisions, a full chronological year, and a contingency set large and sparsely binding enough for screening to pay — where the integer master exercises Theorem 3.3, tailing-off makes stabilisation pay, and a sparse binding set makes screening economical. We present those results in the nested-engine companion work (in preparation) rather than assert them here; doing otherwise would over-read a deliberately stylised instance.

6.5 Reproducibility

The case study is reproducible end-to-end from the Python implementation accompanying this paper (available on request). The random seed is fixed at 20 260 901 throughout (scenario identifier `gb30_dcopf_n1binding_seed20260901`), so all numerical results are reproducible under the recorded code version. The numerical results cited above are read directly from `results/variant_comparison.jso` `results/variant_summary.csv` and `results/price_nse_scope.json`, the certification evidence from `results/level_bundle_cert.json`; the figures are produced from the same run.

7 Discussion

7.1 From the single-level foundation to the nested engine

The foundation of this paper is the degenerate one-block case of the production AFN-B engine. The full engine adds an inner decomposition over chronological storage, so that long horizons with seasonal linkage become tractable; its rigour, its convergence, and its multistage and integer-recourse treatment (the latter along the SDDP/SDDiP route, Pereira and Pinto, 1991; Zou et al., 2019) are specified in the nested-engine companion work (in preparation). The stabilisation studied here (level-bundle master, Magnanti–Wong cuts) and the bound-trap discipline carry over unchanged; this paper is the verified foundation they build on.

7.2 Connection to N-k security

The oracle generalises to N-k security (Wang et al., 2013; Capitanescu et al., 2011) by extending the candidate set $\mathcal{C}$ to combinations of k simultaneously-out elements. The combinatorial blow-up is mitigated by the same screening logic — only combinations that the LODF analysis flags as overloading produce LP solves — so the cost scales with the size of the binding set rather than the full $\binom{|\mathcal{L}|}{k}$ enumeration. As with N-1, the screening *economy* is realised only when the binding set is sparse, which the stylised instance is not constructed to exhibit.

7.3 Limitations

We note four limitations. First, the model uses the linearised DC power-flow model (full DC, with angles, not transport); AC OPF requires the SOCP relaxation (Low, 2014) or convex inner approximations (Coffrin et al., 2016), both of which preserve the decomposition structure but complicate cut generation. Second, the first-stage line builds are LP-relaxed in the demonstrated instance, so the integer-master behaviour of Theorem 3.3 is specified but not exercised numerically here; the integer-recourse caveat of Remark 3.1 likewise means the LP-duality cut is valid only because operations are LP-relaxed. Third, N-1 is cost-internalised (priced corrective redispatch) rather than hard-secured by default; the feasibility-cut variant of Section 2.4 is the route to hard securing. Fourth, the level-bundle master requires a QP solver (CLARABEL here, with OSQP as fallback); the foundation offers no pure-LP variant with equivalent stability, and robust, cheap LP-only stabilisation remains the natural production preference.

8 Conclusion

We have specified and verified the single-level foundation of the CENovaSage AFN-B engine: a two-stage, security-constrained Benders decomposition over a full DC power-flow network, with the cut-validity, bound-trap, and integer-recourse disciplines stated precisely, two stabilisation levers (Magnanti–Wong cuts, a level-bundle QP) studied against the naive single-dual baseline, and N-1 security embedded as priced corrective redispatch through a PTDF/LODF separation oracle. The convergence results are restated at the rigour the foundation requires — finite termination of the mixed-integer master by a finitely-many-dual-bases argument, a total-cut bound for the oracle, and the level-bundle rate stated qualitatively with its strong-convexity hypothesis flagged as non-binding for piecewise-linear recourse. A stylised GB 30-zone instance, engineered so an N-1 boundary bottleneck genuinely binds, demonstrates end-to-end correctness: all six cut/master × security configurations agree on the optimum £74.67M to within the stated tolerance (naive and Magnanti–Wong to machine precision), every power-systems and optimality invariant passes to numerical tolerance, every stored cut is audited as a valid tight minorant, and the oracle reproduces the full-enumeration optimum to machine precision.

We have been deliberate about what this foundation does and does not establish. It is a correctness and mechanics demonstrator. It makes no stabilisation speed-up claim — at this scale the stabilised configurations are in fact slower, the Magnanti–Wong auxiliary solve doubling the naive wall time and the level-bundle master taking seventeen conservative iterations against the naive three — and no screening-economy claim, because the instance is built so that every monitored contingency binds. Those quantitative properties, together with the nested chronological-storage machinery and the integer/multistage recourse treatment, are the province of the production AFN-B engine and its companion study (in preparation). The methodological pieces are individually well-established; the contribution here is a verified, auditable foundation on which the production engine is built, with each claim scoped to what the evidence supports.

Acknowledgements

This work develops methodology first prototyped internally at Compounding Energy Ltd and integrates threads from (Magnanti and Wong, 1981; Lemaréchal et al., 1995; Pereira and Pinto, 1991; Muñoz et al., 2016) alongside contemporary stabilisation work (de Oliveira and Sagastizábal, 2014) and the in-out separation of Bonami et al. (2020). The case study is reproducible from the Python implementation accompanying this paper (available on request). Comments are welcome to `christopher@compoundingenergy.com`.

References

Jacques F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4:238–252, 1962. doi: 10.1007/BF01386316.

Pierre Bonami, Domenico Salvagnin, and Andrea Tramontani. Implementing automatic Benders decomposition in a modern MIP solver. In *Integer Programming and Combinatorial Optimization (IPCO)*, pages 78–90. Springer, 2020.

J. Frédéric Bonnans, J. Charles Gilbert, Claude Lemaréchal, and Claudia A. Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*. Universitext. Springer, 2nd edition, 2006.

Florin Capitanescu, José Luis Martinez Ramos, Patrick Panciatici, Daniel Kirschen, Alejandro Marano Marcolini, Ludovic Platbrood, and Louis Wehenkel. State-of-the-art, challenges, and future trends in security constrained optimal power flow. *Electric Power Systems Research*, 81 (8):1731–1741, 2011. doi: 10.1016/j.epsr.2011.04.003.

C. Coffrin, H. L. Hijazi, and P. Van Hentenryck. The QC relaxation: a theoretical and computational study on optimal power flow. *IEEE Transactions on Power Systems*, 31(4):3008–3018, 2016.

D. L. Cole, M. Lau, X. Dai, S. Chakrabarti, and Jesse D. Jenkins. Analyzing performance and scalability of Benders decomposition for generation and transmission expansion planning models, 2026. arXiv:2603.29867 [math.OC].

Welington de Oliveira and Claudia Sagastizábal. Level bundle methods for oracles with on-demand accuracy. *Optimization Methods and Software*, 29(6):1180–1209, 2014. doi: 10.1080/10556788.2013.871282.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*, volume 306 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1993.

James E. Kelley. The cutting-plane method for solving convex programs. *Journal of the SIAM*, 8(4):703–712, 1960. doi: 10.1137/0108053.

Guanghui Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. *Mathematical Programming*, 149:1–45, 2015. doi: 10.1007/s10107-013-0737-x.

Claude Lemaréchal, Arkadi Nemirovskii, and Yurii Nesterov. New variants of bundle methods. *Mathematical Programming*, 69:111–147, 1995. doi: 10.1007/BF01585555.

S. H. Low. Convex relaxation of optimal power flow—part i: Formulations and equivalence. *IEEE Transactions on Control of Network Systems*, 1(1):15–27, 2014.

T. L. Magnanti and R. T. Wong. Accelerating Benders decomposition: algorithmic enhancement and model selection criteria. *Operations Research*, 29(3):464–484, 1981.

Francisco D. Muñoz, Benjamin F. Hobbs, and Jean-Paul Watson. New bounding and decomposition approaches for MILP investment problems: Multi-area transmission and generation planning under policy constraints. *European Journal of Operational Research*, 248(3):888–898, 2016. doi: 10.1016/j.ejor.2015.07.057.

M. V. F. Pereira and L. M. V. G. Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52:359–375, 1991.

Ragheb Rahmaniani, Teodor Gabriel Crainic, Michel Gendreau, and Walter Rei. The Benders decomposition algorithm: A literature review. *European Journal of Operational Research*, 259 (3):801–817, 2017. doi: 10.1016/j.ejor.2016.12.005.

Andrzej Ruszczyński. A regularized decomposition method for minimizing a sum of polyhedral functions. *Mathematical Programming*, 35:309–333, 1986. doi: 10.1007/BF01580884.

Hanif D. Sherali and Barbara M. P. Fraticelli. A modification of Benders' decomposition algorithm for discrete subproblems: An approach for stochastic programs with integer recourse. *Journal of Global Optimization*, 22:319–342, 2002. doi: 10.1023/A:1013827600901.

Richard M. Van Slyke and Roger J.-B. Wets. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics*, 17(4): 638–663, 1969. doi: 10.1137/0117061.

Qianfan Wang, Jean-Paul Watson, and Yongpei Guan. Two-stage robust optimization for N-k contingency-constrained unit commitment. *IEEE Transactions on Power Systems*, 28(3): 2366–2375, 2013. doi: 10.1109/TPWRS.2013.2244104.

Jikai Zou, Shabbir Ahmed, and Xu Andy Sun. Stochastic dual dynamic integer programming. *Mathematical Programming*, 175:461–502, 2019. doi: 10.1007/s10107-018-1249-5.

A Notation glossary

Symbol	Meaning
$\mathcal{N}, \mathcal{L}, \mathcal{K}, \mathcal{S}, \mathcal{T}, \mathcal{C}$	Sets: zones, lines, technologies, scenarios, time periods, monitored contingencies
$\boldsymbol{x}^{\text{gen}}, \boldsymbol{x}^{\text{line}}$	Investment in generation and transmission (line builds LP-relaxed here, binary in production)
$\boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}, \boldsymbol{r}$	Dispatch: generation, flow, voltage angle, load shedding
$Q_s(\boldsymbol{x})$	Per-scenario base-case dispatch cost-to-go function
$\tilde{Q}_s(\boldsymbol{x})$	Security-constrained recourse: worst case over the base and VOLL-priced post-contingency dispatches, eq. (3)
η_s	Epigraph variable approximating Q_s (θ is reserved for bus voltage angles)
$\alpha_s, \boldsymbol{\beta}_s$	Constant and gradient of a Benders cut
ν^{voll}	Value of lost load (a round 10 000 GBP/MWh, within the range of published GB VoLL estimates)
κ	Level-bundle level fraction (0.6 in the case study)
$\gamma, n_{\max}$	Oracle screen margin (0.9) and per-call post-contingency solve cap (8); case-study settings
Φ, Ψ	PTDF (power-transfer distribution factor) and LODF (line-outage distribution factor) matrices
$\mathcal{C}^{\text{cand}}$	Contingencies screened in by the LODF pre-screen as candidate violations
$\mathcal{V}_{\text{int}}$	Set of distinct integer first-stage points (finite when line builds are binary)

B Level-bundle rate constant

For completeness we record the explicit form of the constant in the $O(1/\sqrt{k})$ rate of Theorem 4.1, with its dimensional reading; we relegate it here rather than to the main text because the binding guarantee for the piecewise-linear recourse of this paper is the qualitative rate, not the constant (Remark 4.2).

Let $f := \langle \boldsymbol{c}, \cdot \rangle + \sum_s p_s Q_s$ be convex and L-Lipschitz (in the norm used by the proximal term) on the master feasible region, whose diameter is D, and let $\kappa \in (0, 1)$ be the level fraction. The level-method analysis of Lemaréchal et al. (1995, Thm. 3.5) gives

$$\mathrm{UB}_k - z^* \;\le\; \frac{D\,L}{\sqrt{2\,\kappa(1-\kappa)\,k}}, \qquad k \ge 1, \tag{8}$$

so the rate constant is $DL/\sqrt{2\kappa(1-\kappa)}$. Dimensionally, D has units of the decision vector (MW for capacities) and L has units of cost per unit decision (GBP per MW), so DL has units of cost (GBP) and sets the scale of the guaranteed per-iteration progress. The factor $1/\sqrt{2\kappa(1-\kappa)}$ is dimensionless, minimised at $\kappa = 1/2$ (value $\sqrt{2}$); at the $\kappa = 0.6$ used here it is ≈ 1.44 (Remark 4.3). The bound is a worst-case asymptotic statement and is uniformly optimal for non-smooth Lipschitz convex problems (Lan, 2015); it does not sharpen to a geometric rate here because f is piecewise-linear and hence not strongly convex (Remark 4.2).

C Reproducibility

The case study reported in Section 6 is reproducible end-to-end from the Python implementation accompanying this paper (`case_study_wp02/simulate_gb30.py`, available on request; the case-study directory retains its original working name). The implementation requires Python 3.10+, NumPy, SciPy (LP solves via the HiGHS backend of `scipy.optimize.linprog`), CVXPY with CLARABEL and OSQP (the level-bundle QP master), Pandas, and Matplotlib. The random seed is fixed at 20 260 901 throughout (scenario identifier `gb30_dcopf_n1binding_seed20260901`); the recorded code version is logged in the `_meta` block of `results/variant_comparison.json`. The numerical values quoted in the text are read directly from `results/variant_comparison.json`, `results/variant_summary.csv` and `results/price_nse_scope.json`; the per-iteration certification evidence for the level-bundle and Magnanti–Wong configurations (QP execution and trial-point movement, Pareto-versus-fallback outcomes, trajectory distinctness, six-way optimum agreement) is written to `results/level_bundle_cert.json`; the figures `convergence_comparison.pdf` and `walltime_comparison.pdf` are written to `figures/` by the same run.