Making Every Bit Count for A-Optimal State Estimation

Cameron Khanpour,   Daniel Turizo,   Samuel Talkington

Abstract— **We study the problem of controlling how a limited communication bandwidth budget is allocated across heterogeneously quantized sensor measurements. The performance criterion is the trace of the error covariance matrix of the linear minimum mean square error (LMMSE) state estimator, i.e., an A-optimal design criterion. Minimizing this criterion with a bit budget constraint yields a nonconvex optimization problem. We derive a formula that reduces each evaluation of the gradient to a single Cholesky factorization. This enables efficient optimization by both a projection-free Frank–Wolfe method (with a computable convergence certificate) and an interior point method with L-BFGS Hessian approximation over the problem's continuous relaxation. A largest remainder rounding procedure recovers integer bit allocations with a bound on the quality of the rounded solution. Numerical experiments in IEEE power grid test cases with up to 300 buses compare both solvers and demonstrate that the analytic gradient is the key computational enabler for both methods. Additionally, the heterogeneous bit allocation is compared to standard uniform bit allocation on the 500 bus IEEE power grid test case.**

I. INTRODUCTION

Analog-to-digital *quantization* methods improve computational and communication efficiency by mapping continuous measurements to discrete intervals [1]. While coarse quantization accelerates computations [2], it introduces *nonlinear*, typically *non-Gaussian*, measurement noise [3], [4], which contrasts with the additive Gaussian noise assumptions that are frequently used for analyzing state estimation algorithms.

At the same time, sensing technologies are increasingly widespread in modern engineering practice. For example, advanced metering infrastructure (AMI) has become widespread in electric power distribution infrastructure, where *variable precision* sensing has emerged as a promising operational paradigm for communication-constrained sensing. To address these challenges, we cast heterogeneous quantizer bit allocation under a global bandwidth budget as an A-optimal design problem for linear minimum mean square error (LMMSE) estimation—i.e., minimizing the trace of the covariance matrix of the error of the LMMSE estimator. The resource allocation acts on measurement precisions and induces a nonconvex optimization problem that falls outside the scope of existing sensor selection methods.

a) Related Work: There are a number of related works. In [5], the discrete sensor placement problem was studied under a D-optimal criterion; the relaxed problem is solved

without guarantees of optimality. Budget-constrained D-optimal design was recently studied in [6]. Similarly, [7] applied sensor placement for water networks; they focus on a class of objective functions that possess other structured properties, such as submodularity (this property is useful to prove theoretical guarantees of greedy based algorithms). While A-optimal design does not have (super)modularity in general, recent research in [8] showed that A-optimal design enjoys *approximate* supermodularity. Combinatorial approximation algorithms for A-optimal design have also been developed [9], [10]. Frank–Wolfe algorithms [11] for the classical A-optimal design problem were analyzed in [12], where the information matrix is linear in the design weights and the problem is convex. The exponential mapping between sensor bitrate and precision in our formulation breaks this linearity and introduces nonconvexity, requiring a different convergence analysis. The work of [13] is also related and shows that the dual of a very similar problem to our own is convex; an earlier paper from the same group [14] proposes a greedy bit assignment heuristic. This line of inquiry enables valuable insights on the energy savings obtained by a chosen bit allocation policy.

Another component of this work is the incorporation of the effects of quantization and communication bandwidth in state estimation. Quantization has a significant effect on the accuracy of state estimation in scenarios limited by bandwidth. In communication-constrained settings—such as remote monitoring over low-bandwidth links or large-scale sensor networks with shared communication channels—the number of bits allocated to each measurement directly determines the quantization noise variance and, consequently, the estimation quality. This connects to the classical theory of optimal experimental design [15], where measurement resources are allocated to minimize estimation error.

Quantized estimation has been addressed in the Kalman filtering [16], [17], sensor network placement [18], and topology learning [19] settings. In power systems, the literature on state estimation has also focused on the importance of heterogeneous data sources, communication constraints, and measurement quality for large scale networks [20].

b) Contributions: In this paper, we study optimal bit allocation for LMMSE state estimation under a global communication budget. We formulate the problem as a nonconvex A-optimal design problem with the following contributions:

1) We present two algorithms: i.) A first-order Frank–Wolfe algorithm with a closed-form linear minimization oracle (Proposition 3), and show that the minimum Frank–Wolfe gap over iterates converges to zero at rate $O(1/\sqrt{T})$ (Theorem 1), providing a computable convergence certificate. ii.) A second-order interior

point algorithm. We derive a closed-form gradient for both FW and the interior point method to reduce each evaluation to a single Cholesky factorization of the information matrix. Further, the interior point method is accelerated with a L-BFGS Hessian approximation and converges in very few iterations.

2) We introduce a solver-agnostic largest remainder rounding procedure (Algorithm 2) that maps continuous relaxations to integer allocations that are guaranteed to be feasible and have bounded solution quality gap.

3) We evaluate the speed and solution quality of the algorithms in the practical setting of experimental design for power network state estimation within different regimes ("sensor rich" and limited bandwidth). We also validate the problem formulation of heterogeneous bit allocation to a standard uniform bit allocation, achieving up to 53% improvement in the limited bandwidth regime.

The remainder of the paper is organized as follows. Section II introduces the measurement model, the LMMSE estimator, and the resulting bit-allocation problem under a global communication budget. Section III presents the proposed solution methods for the relaxed problem, including the analytic gradient, the Frank–Wolfe algorithm and its convergence guarantee, and an interior point approach. Section IV describes the rounding procedure used to recover integer bit allocations from relaxed solutions along with guarantees. Section V reports numerical experiments on IEEE power system test cases, comparing solver performance and evaluating the benefits of heterogeneous allocation relative to uniform allocation. Finally, Section VI concludes the paper and outlines directions for future work.

II. PROBLEM FORMULATION

A. Measurement Model and Estimation

Given a fixed sensing matrix $\boldsymbol{H} \in \mathbb{R}^{m \times d}$ whose rows are $\{\boldsymbol{h}_i^\top\}_{i=1}^m$, where each $\boldsymbol{h}_i \in \mathbb{R}^d$, corresponding to m linear projections of an unknown state vector $\boldsymbol{x} \in \mathbb{R}^d$, we consider heterogeneously quantized measurements with quantization bin widths $\{\Delta_i\}_{i=1}^m$. We assume throughout that $\boldsymbol{h}_i \neq \boldsymbol{0}$ for all i, i.e., every sensor observes a nontrivial linear combination of the state.

Each quantized measurement i is generated from a uniformly dithered quantization function $\mathcal{Q}_i : \mathbb{R} \to \mathbb{R}$ (see [1], [4]) with bin width $\Delta_i > 0$ such that

$$\mathcal{Q}_i(\langle \boldsymbol{h}_i, \boldsymbol{x} \rangle) = \Delta_i \cdot \left(\left\lfloor \frac{\langle \boldsymbol{h}_i, \boldsymbol{x} \rangle + \tau_i}{\Delta_i} \right\rfloor + \frac{1}{2} \right), \quad (1)$$

where $\tau_i \sim \text{Uniform}(-\frac{\Delta_i}{2}, \frac{\Delta_i}{2})$, and $\mathsf{E}_{\tau_i}\left[\mathcal{Q}_i(\langle \boldsymbol{h}_i, \boldsymbol{x} \rangle) \mid \boldsymbol{h}_i \right] = \langle \boldsymbol{h}_i, \boldsymbol{x} \rangle$, i.e., $\mathcal{Q}_i(\langle \boldsymbol{h}_i, \boldsymbol{x} \rangle)$ is conditionally unbiased. The independent dither τ_i is noise purposely applied prior to quantization; see [1], [21]. The quantization error is well approximated by the additive model

$$\boldsymbol{y} := \boldsymbol{H}\boldsymbol{x} + \boldsymbol{z}, \quad (2)$$

where $\mathsf{E}\left[\boldsymbol{z}\boldsymbol{z}^\top\right] = \frac{1}{12}\text{diag}(\Delta_1^2, \dots, \Delta_m^2)$. Set $d_i := \Delta_i^2/12$ and $\boldsymbol{D} := \text{diag}(d_1, \dots, d_m)$. Suppose $\boldsymbol{x}$ is zero-mean with prior covariance $\boldsymbol{C_x} \succ 0$. The LMMSE state estimator is

$$\hat{\boldsymbol{x}} = \boldsymbol{C_x}\boldsymbol{H}^\top \left(\boldsymbol{H}\boldsymbol{C_x}\boldsymbol{H}^\top + \boldsymbol{D} \right)^{-1} \boldsymbol{y}, \quad (3)$$
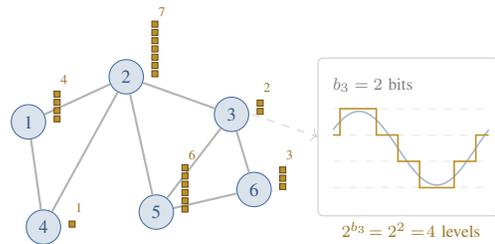
$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{z}, \quad \sum_{i=1}^m b_i \leq B$$

Fig. 1. Bit allocation on a sensor network. Each node i receives b_i bits (■ = 1 bit) subject to budget $\sum_i b_i \leq B$. Inset: 2-bit quantization maps an analog measurement to $2^{b_i} = 4$ discrete levels.

with error covariance

$$\boldsymbol{C_\varepsilon} = \left(\boldsymbol{H}^\top \boldsymbol{D}^{-1} \boldsymbol{H} + \boldsymbol{C_x}^{-1} \right)^{-1}. \quad (4)$$

Quantization error is generally signal-dependent, so an additive noise model can be inaccurate, especially at coarse resolutions [1]. Our use of uniformly dithered quantizers in (1) is intended to mitigate this dependence and justify the covariance model [21]. Accordingly, the optimization problem below should be interpreted as the A-optimal design problem induced by this dithered additive noise LMMSE model.

B. Bandwidth Allocation Problem

We allocate a limited number of quantization bits across m channels to minimize the MSE. For a b_i-bit uniform quantizer with dynamic range R_i,

$$\Delta_i = \frac{R_i}{2^{b_i}}, \quad d_i = \frac{R_i^2}{12 \cdot 4^{b_i}}, \quad \rho_i := d_i^{-1} = \kappa_i \, 4^{b_i}, \quad (5)$$

where $\kappa_i := 12/R_i^2$. Allocating more bits to channel i increases its precision exponentially. The integer program is

$$\min_{\boldsymbol{b} \in \mathbb{Z}_+^m} \quad \text{tr}\big(\boldsymbol{C_\varepsilon}(\boldsymbol{b})\big) \quad \text{s.t.} \quad \boldsymbol{1}^\top \boldsymbol{b} \leq B. \quad (6)$$

The standard continuous relaxation replaces $\mathbb{Z}_+^m$ by $\mathbb{R}_+^m$. Define $\mathcal{B} := \{\boldsymbol{b} \in \mathbb{R}_+^m : \boldsymbol{1}^\top \boldsymbol{b} \leq B\}$ and write

$$\boldsymbol{M}(\boldsymbol{b}) := \boldsymbol{C_x}^{-1} + \boldsymbol{H}^\top \text{diag}(\boldsymbol{\rho}(\boldsymbol{b}))\boldsymbol{H}, \quad \boldsymbol{C_\varepsilon}(\boldsymbol{b}) := \boldsymbol{M}(\boldsymbol{b})^{-1}. \quad (7)$$

The relaxed problem is

$$\min_{\boldsymbol{b} \in \mathcal{B}} \quad F(\boldsymbol{b}) := \text{tr}\big(\boldsymbol{C_\varepsilon}(\boldsymbol{b})\big). \quad (8)$$

The feasible set $\mathcal{B}$ is compact and convex, but F is generally nonconvex in the bit variables due to the exponential reparameterization $\rho_i = \kappa_i 4^{b_i}$. Introducing the auxiliary function $f(\boldsymbol{\rho}) := \text{tr}((\boldsymbol{C_x}^{-1} + \boldsymbol{H}^\top \text{diag}(\boldsymbol{\rho})\boldsymbol{H})^{-1})$ in precision variables, we have $F(\boldsymbol{b}) = f(\boldsymbol{\rho}(\boldsymbol{b}))$. While f is convex in $\boldsymbol{\rho}$, the chain rule yields

$$\nabla^2 F(\boldsymbol{b}) = (\ln 4)^2 \Big[\text{diag}(\boldsymbol{\rho})\nabla^2 f(\boldsymbol{\rho}) \, \text{diag}(\boldsymbol{\rho}) \\ + \text{diag}\big(\boldsymbol{\rho} \odot \nabla f(\boldsymbol{\rho})\big) \Big], \quad (9)$$

where the second term is negative semidefinite (since $\nabla f \leq \boldsymbol{0}$ componentwise), so convexity does not transfer from $\boldsymbol{\rho}$ to $\boldsymbol{b}$.

Remark (Convexity in precision space). Although $F(\boldsymbol{b})$ is nonconvex in the bit variables, the function $f(\boldsymbol{\rho}) = \operatorname{tr}((\boldsymbol{C}_{\boldsymbol{x}}^{-1} + \boldsymbol{H}^{\top} \operatorname{diag}(\boldsymbol{\rho})\boldsymbol{H})^{-1})$ is convex in the precision variables $\boldsymbol{\rho}$. However, the budget constraint $\mathbf{1}^{\top}\boldsymbol{b} \leq B$ becomes $\sum_{i=1}^{m} \log_4(\rho_i/\kappa_i) \leq B$ in $\boldsymbol{\rho}$-space, which is *not* convex (it is the sublevel set of a concave function). Thus, the $\boldsymbol{b}$-space formulation trades a nonconvex objective for a convex (polyhedral) feasible set, which is precisely what enables the Frank–Wolfe method with a closed-form linear minimization oracle. This tradeoff is complementary to the dual formulation of [13], which achieves a convex objective at the cost of a more complex constraint geometry.

Lemma 1 (Gradient with respect to bits). *For each measurement $i = 1, \ldots, m$,*

$$\frac{\partial F}{\partial b_i}(\boldsymbol{b}) = -(\ln 4)\,\rho_i(\boldsymbol{b})\,\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}(\boldsymbol{b})^2\boldsymbol{h}_i. \tag{10}$$

Equivalently,

$$\nabla F(\boldsymbol{b}) = -(\ln 4)\,\boldsymbol{\rho}(\boldsymbol{b}) \odot \left[\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}(\boldsymbol{b})^2\boldsymbol{h}_i\right]_{i=1}^{m}. \tag{11}$$

In particular, $\partial F/\partial b_i < 0$ whenever $\boldsymbol{h}_i \neq \boldsymbol{0}$. Each gradient evaluation reduces to a single Cholesky factorization of $\boldsymbol{M}(\boldsymbol{b})$: the diagonal entries $\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_i$ are extracted column-by-column from $\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{H}^{\top}$ without forming the full $m \times m$ product, at cost $O(d^2m)$ dominated by the $O(d^3)$ factorization when $m = O(d)$.

Proof: Define

$$\boldsymbol{A}(\boldsymbol{\rho}) := \boldsymbol{C}_{\boldsymbol{x}}^{-1} + \boldsymbol{H}^{\top}\operatorname{diag}(\boldsymbol{\rho})\boldsymbol{H}, \quad \boldsymbol{C}_{\boldsymbol{\varepsilon}}(\boldsymbol{\rho}) = \boldsymbol{A}(\boldsymbol{\rho})^{-1}.$$

Since $\partial\boldsymbol{A}/\partial\rho_i = \boldsymbol{H}^{\top}\boldsymbol{e}_i\boldsymbol{e}_i^{\top}\boldsymbol{H} = \boldsymbol{h}_i\boldsymbol{h}_i^{\top}$, the derivative identity for matrix inverses gives

$$\frac{\partial\boldsymbol{C}_{\boldsymbol{\varepsilon}}}{\partial\rho_i} = -\boldsymbol{C}_{\boldsymbol{\varepsilon}}\frac{\partial\boldsymbol{A}}{\partial\rho_i}\boldsymbol{C}_{\boldsymbol{\varepsilon}} = -\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_i\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}.$$

Taking traces,

$$\frac{\partial f}{\partial\rho_i} = \operatorname{tr}\left(\frac{\partial\boldsymbol{C}_{\boldsymbol{\varepsilon}}}{\partial\rho_i}\right) = -\operatorname{tr}\left(\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_i\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\right) = -\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_i.$$

Since $\boldsymbol{C}_{\boldsymbol{\varepsilon}} \succ 0$, the quadratic form $\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_i$ is nonnegative. Now $F(\boldsymbol{b}) = f(\boldsymbol{\rho}(\boldsymbol{b}))$ with $\rho_i(\boldsymbol{b}) = \kappa_i 4^{b_i}$, so by chain rule

$$\frac{\partial F}{\partial b_i} = \frac{\partial f}{\partial\rho_i} \cdot \frac{\partial\rho_i}{\partial b_i} = \left(-\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}(\boldsymbol{b})^2\boldsymbol{h}_i\right) \cdot (\ln 4)\rho_i(\boldsymbol{b}),$$

which yields (10). The vector form (11) follows by stacking. If $\boldsymbol{h}_i \neq \boldsymbol{0}$, then $\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}(\boldsymbol{b})^2\boldsymbol{h}_i > 0$ because $\boldsymbol{C}_{\boldsymbol{\varepsilon}}(\boldsymbol{b}) \succ 0$, so the derivative is strictly negative. ∎

Proposition 1 (Budget saturation). *If $\boldsymbol{h}_i \neq \boldsymbol{0}$ for all i and $\boldsymbol{b}^{\star}$ minimizes (8), then $\mathbf{1}^{\top}\boldsymbol{b}^{\star} = B$.*

Proof: Suppose for contradiction that $\mathbf{1}^{\top}\boldsymbol{b}^{\star} < B$. Then there exists at least one index i such that $\boldsymbol{b}^{\star} + t\boldsymbol{e}_i \in \mathcal{B}$ for all sufficiently small $t > 0$. By Lemma 1, $\partial F/\partial b_i(\boldsymbol{b}^{\star}) < 0$, so increasing b_i decreases the objective, contradicting optimality of $\boldsymbol{b}^{\star}$. ∎

Proposition 2 (Lipschitz continuity of the gradient). *The function F is C^{∞} on $\mathbb{R}^m$, and ∇F is L-Lipschitz on $\mathcal{B}$ with*

$$L = (\ln 4)^2\,\|\boldsymbol{C}_{\boldsymbol{x}}\|_2\,(2m + 1). \tag{12}$$

Proof: Recall that $F(\boldsymbol{b}) = f(\boldsymbol{\rho}(\boldsymbol{b}))$ and $\rho_i(\boldsymbol{b}) = \kappa_i 4^{b_i}$, where

$$f(\boldsymbol{\rho}) = \operatorname{tr}\left((\boldsymbol{C}_{\boldsymbol{x}}^{-1} + \boldsymbol{H}^{\top}\operatorname{diag}(\boldsymbol{\rho})\boldsymbol{H})^{-1}\right).$$

Since each coordinate map $\boldsymbol{b} \mapsto \rho_i(\boldsymbol{b}) = \kappa_i 4^{b_i}$ is smooth and strictly positive on $\mathbb{R}^m$, and matrix inversion is smooth on the positive definite cone, F is C^{∞} on $\mathbb{R}^m$. Since $\boldsymbol{C}_{\boldsymbol{\varepsilon}}(\boldsymbol{b}) \succ 0$ for all $\boldsymbol{b}$, and $\boldsymbol{C}_{\boldsymbol{\varepsilon}}(\boldsymbol{b}) \preceq \boldsymbol{C}_{\boldsymbol{x}}$ implies $\|\boldsymbol{C}_{\boldsymbol{\varepsilon}}(\boldsymbol{b})\|_2 \leq \|\boldsymbol{C}_{\boldsymbol{x}}\|_2$. From Lemma 1, $\frac{\partial f}{\partial\rho_i}(\boldsymbol{\rho}) = -\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_i$. Differentiating once more with respect to ρ_j gives

$$\frac{\partial^2 f}{\partial\rho_j\partial\rho_i}(\boldsymbol{\rho}) = 2\operatorname{tr}\left(\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_i\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_j\boldsymbol{h}_j^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\right). \tag{13}$$

Fix i, and write $\boldsymbol{A}_i := \boldsymbol{C}_{\boldsymbol{x}}^{-1} + \sum_{k\neq i}\rho_k\boldsymbol{h}_k\boldsymbol{h}_k^{\top} \succ 0$. By the Sherman–Morrison–Woodbury identity,

$$\boldsymbol{C}_{\boldsymbol{\varepsilon}} = \boldsymbol{A}_i^{-1} - \frac{\rho_i\,\boldsymbol{A}_i^{-1}\boldsymbol{h}_i\boldsymbol{h}_i^{\top}\boldsymbol{A}_i^{-1}}{1 + \rho_i\boldsymbol{h}_i^{\top}\boldsymbol{A}_i^{-1}\boldsymbol{h}_i}.$$

Hence, with $\boldsymbol{h}_i^{\top}\boldsymbol{A}_i^{-1}\boldsymbol{h}_i \geq 0$, we obtain

$$\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_i = \frac{\boldsymbol{h}_i^{\top}\boldsymbol{A}_i^{-1}\boldsymbol{h}_i}{1 + \rho_i\boldsymbol{h}_i^{\top}\boldsymbol{A}_i^{-1}\boldsymbol{h}_i} \leq \frac{1}{\rho_i}.$$

Therefore

$$\rho_i\,\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_i \leq 1. \tag{14}$$

Since $\boldsymbol{C}_{\boldsymbol{\varepsilon}} \succeq 0$ and $\|\boldsymbol{C}_{\boldsymbol{\varepsilon}}\|_2 \leq \|\boldsymbol{C}_{\boldsymbol{x}}\|_2$, we have

$$\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2 \preceq \|\boldsymbol{C}_{\boldsymbol{\varepsilon}}\|_2\,\boldsymbol{C}_{\boldsymbol{\varepsilon}} \preceq \|\boldsymbol{C}_{\boldsymbol{x}}\|_2\boldsymbol{C}_{\boldsymbol{\varepsilon}}.$$

Thus $\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_i \leq \|\boldsymbol{C}_{\boldsymbol{x}}\|_2\,\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_i$. Using (14),

$$\rho_i\left|\frac{\partial f}{\partial\rho_i}\right| = \rho_i\,\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_i \leq \|\boldsymbol{C}_{\boldsymbol{x}}\|_2\,\rho_i\,\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_i \leq \|\boldsymbol{C}_{\boldsymbol{x}}\|_2.$$

Therefore

$$\rho_i\left|\frac{\partial f}{\partial\rho_i}\right| \leq \|\boldsymbol{C}_{\boldsymbol{x}}\|_2. \tag{15}$$

From (13),

$$\frac{\partial^2 f}{\partial\rho_j\partial\rho_i} = 2\,(\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_j)(\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_j).$$

Since $\boldsymbol{C}_{\boldsymbol{\varepsilon}} \succeq 0$, the Cauchy–Schwarz inequality in the $\boldsymbol{C}_{\boldsymbol{\varepsilon}}$-inner product yields

$$|\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_j| \leq \left(\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_i\right)^{1/2}\left(\boldsymbol{h}_j^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_j\right)^{1/2} \leq \frac{1}{\sqrt{\rho_i\rho_j}},$$

where the last step uses (14). Since $\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2 \preceq \|\boldsymbol{C}_{\boldsymbol{x}}\|_2\boldsymbol{C}_{\boldsymbol{\varepsilon}}$,

$$\begin{aligned}
|\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_j| &\leq \left(\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_i\right)^{1/2}\left(\boldsymbol{h}_j^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}^2\boldsymbol{h}_j\right)^{1/2} \\
&\leq \|\boldsymbol{C}_{\boldsymbol{x}}\|_2\left(\boldsymbol{h}_i^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_i\right)^{1/2}\left(\boldsymbol{h}_j^{\top}\boldsymbol{C}_{\boldsymbol{\varepsilon}}\boldsymbol{h}_j\right)^{1/2} \\
&\leq \frac{\|\boldsymbol{C}_{\boldsymbol{x}}\|_2}{\sqrt{\rho_i\rho_j}}.
\end{aligned}$$

Combining the two bounds,

$$\left|\frac{\partial^2 f}{\partial\rho_j\partial\rho_i}\right| \leq \frac{2\|\boldsymbol{C}_{\boldsymbol{x}}\|_2}{\rho_i\rho_j},$$

and therefore

$$\rho_i \rho_j \left| \frac{\partial^2 f}{\partial \rho_j \partial \rho_i} \right| \leq 2\|\boldsymbol{C_x}\|_2. \tag{16}$$

From the chain rule with δ_{ij} denoting the Kronecker delta,

$$\frac{\partial^2 F}{\partial b_j \partial b_i} = (\ln 4)^2 \left[\delta_{ij}\, \rho_i \frac{\partial f}{\partial \rho_i} + \rho_i \rho_j \frac{\partial^2 f}{\partial \rho_j \partial \rho_i} \right]. \tag{17}$$

Using (15) and (16),

$$\left| \frac{\partial^2 F}{\partial b_j \partial b_i} \right| \leq (\ln 4)^2 \left[\delta_{ij}\|\boldsymbol{C_x}\|_2 + 2\|\boldsymbol{C_x}\|_2 \right].$$

Hence, for each fixed i,

$$\sum_{j=1}^{m} \left| \frac{\partial^2 F}{\partial b_j \partial b_i} \right| \leq (\ln 4)^2 \sum_{j=1}^{m} \left[\delta_{ij}\|\boldsymbol{C_x}\|_2 + 2\|\boldsymbol{C_x}\|_2 \right]$$
$$= (\ln 4)^2 \|\boldsymbol{C_x}\|_2 (2m+1) =: L.$$

Thus $\|\nabla^2 F(\boldsymbol{b})\|_2 \leq \|\nabla^2 F(\boldsymbol{b})\|_\infty \leq (\ln 4)^2 \|\boldsymbol{C_x}\|_2 (2m+1)$ for all $\boldsymbol{b} \in \mathcal{B}$. Since the Hessian is uniformly bounded on the convex set $\mathcal{B}$, the mean value theorem implies that ∇F is L-Lipschitz continuous on $\mathcal{B}$. ∎

III. BIT ALLOCATION PROCEDURE

We solve the relaxed problem (8) using the Frank–Wolfe (FW) method [11]; see [22], [23] for introductions. Unlike the convex setting of [12], our objective is nonconvex in $\boldsymbol{b}$, so we apply the nonconvex FW analysis of [24]. FW is well suited because the feasible set is compact and convex, the objective is smooth, and the linear minimization oracle (LMO) admits a closed-form solution.

A. Gradient and Linear Minimization Oracle

Given a feasible iterate $\boldsymbol{b} \in \mathcal{B}$, the FW linear minimization oracle solves

$$\boldsymbol{s}(\boldsymbol{b}) \in \arg\min_{\boldsymbol{s} \in \mathcal{B}} \langle \nabla F(\boldsymbol{b}), \boldsymbol{s} \rangle. \tag{18}$$

The FW direction is $\boldsymbol{d}(\boldsymbol{b}) := \boldsymbol{s}(\boldsymbol{b}) - \boldsymbol{b}$, and the FW gap is

$$g_{\mathrm{FW}}(\boldsymbol{b}) := \max_{\boldsymbol{s} \in \mathcal{B}} \langle \boldsymbol{b} - \boldsymbol{s}, \nabla F(\boldsymbol{b}) \rangle = - \langle \boldsymbol{d}(\boldsymbol{b}), \nabla F(\boldsymbol{b}) \rangle. \tag{19}$$

As usual, $g_{\mathrm{FW}}(\boldsymbol{b}) \geq 0$, with equality if and only if $\boldsymbol{b}$ is a first-order stationary point.

Proposition 3 (Closed-form LMO). *Let $\boldsymbol{g} := \nabla F(\boldsymbol{b}) \in \mathbb{R}^m$. Then an optimal solution of (18) is*

$$\boldsymbol{s}^\star = \begin{cases} B\, \boldsymbol{e}_{i^\star}, & \text{if } \min_i g_i < 0 \text{ and } i^\star \in \arg\min_i g_i, \\ \boldsymbol{0}, & \text{if } g_i \geq 0 \text{ for all } i. \end{cases} \tag{20}$$

Under the standing assumption $\boldsymbol{h}_i \neq \boldsymbol{0}$ for all i (so $\boldsymbol{g} < \boldsymbol{0}$ by Lemma 1), this simplifies to

$$\boldsymbol{s}^\star = B\, \boldsymbol{e}_{i^\star}, \quad i^\star \in \arg\min_{1 \leq i \leq m} [\nabla F(\boldsymbol{b})]_i. \tag{21}$$

Proof: The feasible set $\mathcal{B}$ is a polytope with extreme points $\boldsymbol{0}, B\boldsymbol{e}_1, \ldots, B\boldsymbol{e}_m$. Since $\boldsymbol{s} \mapsto \langle \boldsymbol{g}, \boldsymbol{s} \rangle$ is linear, an optimum is attained at a vertex. Evaluating gives $\langle \boldsymbol{g}, \boldsymbol{0} \rangle = 0$ and $\langle \boldsymbol{g}, B\boldsymbol{e}_i \rangle = Bg_i$, so the minimum is $\min\{0, B\min_i g_i\}$, yielding (20). Simplification (21) follows from Lemma 1. ∎

$$g_{\mathrm{FW}}(\boldsymbol{b}^{(t)}) = \left\langle \boldsymbol{b}^{(t)} - \boldsymbol{s}^{(t)}, \nabla F(\boldsymbol{b}^{(t)}) \right\rangle$$
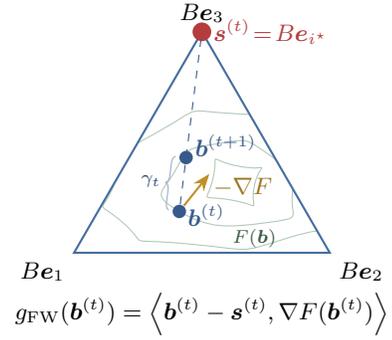
Fig. 2. Frank–Wolfe iteration on the budget simplex $\mathcal{B}$ for $m = 3$. The LMO selects vertex $\boldsymbol{s}^{(t)}$; the next iterate lies on $[\boldsymbol{b}^{(t)}, \boldsymbol{s}^{(t)}]$ with step γ_t.

Corollary 3.1 (Explicit Frank–Wolfe gap). *Under the assumption $\boldsymbol{h}_i \neq \boldsymbol{0}$ for all i,*

$$g_{\mathrm{FW}}(\boldsymbol{b}) = \langle \boldsymbol{b}, \nabla F(\boldsymbol{b}) \rangle - B \min_{1 \leq i \leq m} [\nabla F(\boldsymbol{b})]_i.$$

Proof: By definition of FW gap in (19), $g_{\mathrm{FW}}(\boldsymbol{b}) = \langle \boldsymbol{b} - \boldsymbol{s}(\boldsymbol{b}), \nabla F(\boldsymbol{b}) \rangle$ where $\boldsymbol{s}(\boldsymbol{b}) \in \arg\min_{\boldsymbol{s} \in \mathcal{B}} \langle \nabla F(\boldsymbol{b}), \boldsymbol{s} \rangle$. Since $\nabla F(\boldsymbol{b}) < \boldsymbol{0}$ componentwise by Lemma 1, Proposition 3 gives $\boldsymbol{s}(\boldsymbol{b}) = B\boldsymbol{e}_{i^\star}$ with $i^\star \in \arg\min_i[\nabla F(\boldsymbol{b})]_i$. Hence $g_{\mathrm{FW}}(\boldsymbol{b}) = \langle \boldsymbol{b}, \nabla F(\boldsymbol{b}) \rangle - B\min_i[\nabla F(\boldsymbol{b})]_i$. ∎

B. Convergence Guarantee

Starting from $\boldsymbol{b}^{(0)} \in \mathcal{B}$, the FW method generates

$$\boldsymbol{s}^{(t)} \in \arg\min_{\boldsymbol{s} \in \mathcal{B}} \left\langle \nabla F(\boldsymbol{b}^{(t)}), \boldsymbol{s} \right\rangle, \quad \boldsymbol{d}^{(t)} := \boldsymbol{s}^{(t)} - \boldsymbol{b}^{(t)},$$

and updates $\boldsymbol{b}^{(t+1)} = (1 - \gamma_t)\boldsymbol{b}^{(t)} + \gamma_t \boldsymbol{s}^{(t)}$ with step size

$$\gamma_t = \min \left\{ \frac{g_{\mathrm{FW}}(\boldsymbol{b}^{(t)})}{2LB^2}, 1 \right\}. \tag{22}$$

Since $\mathcal{B}$ is convex, every iterate remains feasible. The step size (22) uses $\mathrm{diam}(\mathcal{B}) = \sqrt{2}B$ by property of simplices.

Theorem 1 (FW convergence for relaxed bit allocation [24]). *Assume $\boldsymbol{h}_i \neq \boldsymbol{0}$ for all i, and let $\{\boldsymbol{b}^{(t)}\}_{t \geq 0}$ be generated by (22) starting from any $\boldsymbol{b}^{(0)} \in \mathcal{B}$. Define $h_0 := F(\boldsymbol{b}^{(0)}) - \min_{\boldsymbol{b} \in \mathcal{B}} F(\boldsymbol{b})$. Then*

$$\min_{0 \leq t \leq T} g_{\mathrm{FW}}(\boldsymbol{b}^{(t)}) \leq \frac{\max\{2h_0,\, 2LB^2\}}{\sqrt{T+1}}. \tag{23}$$

To achieve $g_{\mathrm{FW}}(\boldsymbol{b}^{(t)}) \leq \varepsilon$, it suffices to perform $T + 1 \geq \max\{2h_0, 2LB^2\}^2 / \varepsilon^2$ iterations.

Proof: By Proposition 2, F is continuously differentiable on $\mathbb{R}^m$ with L-Lipschitz gradient on the compact convex set $\mathcal{B}$. The FW curvature constant C_F over $\mathcal{B}$ satisfies

$$C_F \leq L\, \mathrm{diam}(\mathcal{B})^2 = 2LB^2,$$

where $\mathrm{diam}(\mathcal{B}) = \sqrt{2}B$ follows from the vertex structure. Applying [24, Theorem 1] with $C := 2LB^2$ gives

$$\min_{0 \leq t \leq T} g_{\mathrm{FW}}(\boldsymbol{b}^{(t)}) \leq \frac{\max\{2h_0,\, C\}}{\sqrt{T+1}} = \frac{\max\{2h_0,\, 2LB^2\}}{\sqrt{T+1}},$$

which is (23). Rearranging yields the iteration complexity. ∎

Input: $\boldsymbol{H}$, $\boldsymbol{C_x}$, $\boldsymbol{\kappa}$, B, L, ε, $T_{\max}$
Output: feasible $\boldsymbol{b}^{(t)} \in \mathcal{B}$
$\boldsymbol{b}^{(0)} \leftarrow \boldsymbol{0}$
for $t = 0, 1, \dots, T_{\max}$ **do**
$\quad \boldsymbol{\rho}^{(t)} \leftarrow \boldsymbol{\kappa} \odot 4^{\boldsymbol{b}^{(t)}}$
$\quad \boldsymbol{M}^{(t)} \leftarrow \boldsymbol{C_x}^{-1} + \boldsymbol{H}^{\top} \operatorname{diag}(\boldsymbol{\rho}^{(t)}) \boldsymbol{H}$
$\quad \boldsymbol{C_\varepsilon}^{(t)} \leftarrow \left(\boldsymbol{M}^{(t)}\right)^{-1}$
$\quad \boldsymbol{g}^{(t)} \leftarrow -(\ln 4) \boldsymbol{\rho}^{(t)} \odot \operatorname{diag}\left(\boldsymbol{H}(\boldsymbol{C_\varepsilon}^{(t)})^2 \boldsymbol{H}^{\top}\right)$
$\quad i_t \leftarrow \arg\min_{1 \le i \le m} [\boldsymbol{g}^{(t)}]_i$
$\quad \boldsymbol{s}^{(t)} \leftarrow B \, \boldsymbol{e}_{i_t}$
$\quad g_t \leftarrow \left\langle \boldsymbol{b}^{(t)} - \boldsymbol{s}^{(t)}, \boldsymbol{g}^{(t)} \right\rangle$
$\quad$ **if** $g_t \le \varepsilon$ **then return** $\boldsymbol{b}^{(t)}$
$\quad \gamma_t \leftarrow \min\left\{ \frac{g_t}{2LB^2}, 1 \right\}$
$\quad \boldsymbol{b}^{(t+1)} \leftarrow (1 - \gamma_t) \boldsymbol{b}^{(t)} + \gamma_t \boldsymbol{s}^{(t)}$
end
return $\boldsymbol{b}^{(T_{\max}+1)}$
Algorithm 1: Frank–Wolfe for relaxed bit allocation

Input: continuous solution $\bar{\boldsymbol{b}} \in \mathcal{B}$ with $\boldsymbol{1}^{\top} \bar{\boldsymbol{b}} = B$
Output: integral $\hat{\boldsymbol{b}} \in \mathbb{Z}_+^m$ with $\boldsymbol{1}^{\top} \hat{\boldsymbol{b}} = B$
$\boldsymbol{b}^{(0)} \leftarrow \lfloor \bar{\boldsymbol{b}} \rfloor$
$\boldsymbol{r} \leftarrow \bar{\boldsymbol{b}} - \lfloor \bar{\boldsymbol{b}} \rfloor$
$R_{\text{rem}} \leftarrow B - \boldsymbol{1}^{\top} \boldsymbol{b}^{(0)}$
$\hat{\boldsymbol{b}} \leftarrow \boldsymbol{b}^{(0)}$
find a set $S \subseteq \{1, \dots, m\}$ with $|S| = R_{\text{rem}}$ such that

$$S \in \arg\max_{\substack{T \subseteq \{1, \dots, m\} \\ |T| = R_{\text{rem}}}} \sum_{i \in T} r_i$$

(i.e., S indexes the R_{rem} largest components of $\boldsymbol{r}$)
foreach $i \in S$ **do**
$\quad \hat{b}_i \leftarrow \hat{b}_i + 1$
end
return $\hat{\boldsymbol{b}}$
Algorithm 2: Largest remainder rounding procedure

Remark. Theorem 1 guarantees convergence to a first-order stationary point at rate $O(\varepsilon^{-2})$. The stationarity measure is the computable FW gap, obtained at no extra cost once the LMO is solved.

Remark (Adaptive step size). In practice, the global bound L can be conservative. One may replace the step size in Algorithm 1 with an adaptive Lipschitz search: initialize $\hat{L}_0 = 1$; at each iteration, halve $\hat{L}$, then double until $F(\boldsymbol{b} + \gamma_t \boldsymbol{d}^{(t)}) \le F(\boldsymbol{b}^{(t)}) - \gamma_t g_{\text{FW}}(\boldsymbol{b}^{(t)})/2$, capping at L. Since $\hat{L}_t \le L$, Theorem 1 still applies.

C. Interior Point Method with Analytic Gradient

The gradient formula from Lemma 1 can also be supplied to general purpose nonlinear programming solvers. We use the interior point solver Ipopt [25] via JuMP [26] with the option `hessian_approximation = "limited-memory"`, which activates an L-BFGS Hessian approximation. Ipopt replaces the constrained problem (8) with a sequence of log-barrier subproblems

$$\min_{\boldsymbol{b}} \; F(\boldsymbol{b}) - \mu \sum_{i=1}^{m} \ln b_i - \mu \ln(B - \boldsymbol{1}^{\top} \boldsymbol{b}),$$

where $\mu > 0$ is driven to zero. Each subproblem is solved by a damped Newton method on the primal-dual KKT system. Since the constraint structure is simple (m bound constraints plus one linear budget constraint), the KKT system is inexpensive to solve once the gradient is available.

The Hessian $\nabla^2 F(\boldsymbol{b})$ is costly to compute exactly. Instead, we use the limited-memory BFGS (L-BFGS) approximation, which maintains a low-rank estimate from recent gradient differences. With L-BFGS, each iteration requires only the gradient $\nabla F(\boldsymbol{b})$ from Lemma 1—the same $O(d^3)$ Cholesky factorization used by Frank–Wolfe—plus $O(m)$ work for the L-BFGS update and KKT solve.

Remark (Comparison of per-iteration costs). Both Frank–Wolfe and the interior point method are bottlenecked by the same $O(d^3)$ Cholesky factorization for the gradient. FW adds an $O(m)$ LMO step, whereas Ipopt adds an $O(m)$ KKT solve. The key difference is iteration count; while FW and

interior point methods both globally converge sublinearly at rate $O(1/\sqrt{T})$, IPMs also exhibit local superlinear convergence, typically requiring 20–50 iterations. FW provides a computable convergence certificate (the FW gap), while Ipopt reports KKT residuals.

IV. ROUNDING PROCEDURE

The relaxed problem (8) returns a continuous solution $\bar{\boldsymbol{b}} \in \mathcal{B}$, whereas the original problem (6) requires $\boldsymbol{b} \in \mathbb{Z}_+^m$. We compute

$$\boldsymbol{r} := \bar{\boldsymbol{b}} - \lfloor \bar{\boldsymbol{b}} \rfloor \in [0, 1)^m, \quad R_{\text{rem}} := \boldsymbol{1}^{\top} \boldsymbol{r},$$

where $\lfloor \cdot \rfloor$ is computed componentwise. Since $\boldsymbol{1}^{\top} \bar{\boldsymbol{b}} = B$ by Proposition 1 and $\boldsymbol{1}^{\top} \lfloor \bar{\boldsymbol{b}} \rfloor \in \mathbb{Z}_+$, the residual budget is an integer satisfying $0 \le R_{\text{rem}} \le m - 1$.

Similarly to the Quota Method for apportionment [27] and the rounding procedures for approximate experimental designs [28], we use *largest remainder rounding* to round up the R_{rem} largest components of $\boldsymbol{r}$ and round the rest down. Equivalently,

$$\boldsymbol{\xi} \in \arg\max_{\substack{\boldsymbol{\xi} \in \{0,1\}^m \\ \boldsymbol{1}^{\top} \boldsymbol{\xi} = R_{\text{rem}}}} \langle \boldsymbol{r}, \boldsymbol{\xi} \rangle, \quad \hat{\boldsymbol{b}} := \lfloor \bar{\boldsymbol{b}} \rfloor + \boldsymbol{\xi}. \quad (24)$$

Proposition 4 (Feasibility and nearest point property of largest remainder rounding [27]). *Let $\bar{\boldsymbol{b}} \in \mathcal{B}$ satisfy $\boldsymbol{1}^{\top} \bar{\boldsymbol{b}} = B$, and define $\boldsymbol{r} = \bar{\boldsymbol{b}} - \lfloor \bar{\boldsymbol{b}} \rfloor$ and $R_{\text{rem}} = \boldsymbol{1}^{\top} \boldsymbol{r}$. Then $\hat{\boldsymbol{b}}$ defined by (24) belongs to $\mathbb{Z}_+^m$, satisfies $\boldsymbol{1}^{\top} \hat{\boldsymbol{b}} = B$, and solves*

$$\hat{\boldsymbol{b}} \in \arg\min_{\substack{\hat{\boldsymbol{b}} \in \mathbb{Z}_+^m, \\ \hat{\boldsymbol{b}} = \lfloor \bar{\boldsymbol{b}} \rfloor + \boldsymbol{\xi}}} \|\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}\|_2^2 \; \text{s.t.} \; \boldsymbol{\xi} \in \{0,1\}^m, \; \boldsymbol{1}^{\top} \boldsymbol{\xi} = R_{\text{rem}}. \quad (25)$$

Moreover,

$$\|\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}\|_2^2 \le \sum_{i=1}^{m} r_i(1 - r_i). \quad (26)$$

Proof: By construction, $\boldsymbol{\xi} \in \{0,1\}^m$ and $\boldsymbol{1}^{\top} \boldsymbol{\xi} = R_{\text{rem}}$, so $\hat{\boldsymbol{b}} = \lfloor \bar{\boldsymbol{b}} \rfloor + \boldsymbol{\xi} \in \mathbb{Z}_+^m$ and

$$\boldsymbol{1}^{\top} \hat{\boldsymbol{b}} = \boldsymbol{1}^{\top} \lfloor \bar{\boldsymbol{b}} \rfloor + \boldsymbol{1}^{\top} \boldsymbol{\xi} = (B - R_{\text{rem}}) + R_{\text{rem}} = B.$$

Now let $\hat{\boldsymbol{b}} = \lfloor \bar{\boldsymbol{b}} \rfloor + \boldsymbol{\xi}$ with $\boldsymbol{\xi} \in \{0,1\}^m$ and $\boldsymbol{1}^\top \boldsymbol{\xi} = R_{\text{rem}}$. Since $\bar{\boldsymbol{b}} = \lfloor \bar{\boldsymbol{b}} \rfloor + \boldsymbol{r}$, we obtain

$$\|\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}\|_2^2 = \|\boldsymbol{\xi} - \boldsymbol{r}\|_2^2 = \|\boldsymbol{r}\|_2^2 + \|\boldsymbol{\xi}\|_2^2 - 2\langle \boldsymbol{r}, \boldsymbol{\xi} \rangle.$$

Because $\boldsymbol{\xi} \in \{0,1\}^m$ with $\boldsymbol{1}^\top \boldsymbol{\xi} = R_{\text{rem}}$, we have $\|\boldsymbol{\xi}\|_2^2 = R_{\text{rem}}$. Hence

$$\|\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}\|_2^2 = \|\boldsymbol{r}\|_2^2 + R_{\text{rem}} - 2\langle \boldsymbol{r}, \boldsymbol{\xi} \rangle.$$

Therefore minimizing $\|\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}\|_2^2$ over all feasible $\boldsymbol{\xi}$ is equivalent to maximizing $\langle \boldsymbol{r}, \boldsymbol{\xi} \rangle$, which proves (25).

For (26), note that $\boldsymbol{r} \in [0,1]^m$ and $\boldsymbol{1}^\top \boldsymbol{r} = R_{\text{rem}}$, so $\boldsymbol{r}$ is feasible for the continuous relaxation of the maximization problem in (24). Hence $\langle \boldsymbol{r}, \boldsymbol{\xi} \rangle \geq \langle \boldsymbol{r}, \boldsymbol{r} \rangle = \|\boldsymbol{r}\|_2^2$. Substituting into the norm identity above gives

$$\|\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}\|_2^2 = \|\boldsymbol{r}\|_2^2 + R_{\text{rem}} - 2\langle \boldsymbol{r}, \boldsymbol{\xi} \rangle$$
$$\leq R_{\text{rem}} - \|\boldsymbol{r}\|_2^2 = \sum_{i=1}^m r_i(1 - r_i),$$

as claimed. ∎

Theorem 2 (Rounding gap for largest remainder rounding). *Assume $\boldsymbol{h}_i \neq \boldsymbol{0}$ for all i, and let $\bar{\boldsymbol{b}} \in \mathcal{B}$ satisfy $\boldsymbol{1}^\top \bar{\boldsymbol{b}} = B$. Suppose $\bar{\boldsymbol{b}}$ is a KKT point of the relaxed problem (8). Then the largest remainder rounding $\hat{\boldsymbol{b}}$ defined by (24) satisfies*

$$F(\hat{\boldsymbol{b}}) \leq F(\bar{\boldsymbol{b}}) + \frac{L}{2} \sum_{i=1}^m r_i(1 - r_i). \quad (27)$$

In particular,

$$F(\hat{\boldsymbol{b}}) \leq F(\bar{\boldsymbol{b}}) + \frac{L}{2} \min\{R_{\text{rem}}, m/4\}. \quad (28)$$

Proof: Since $\bar{\boldsymbol{b}}, \hat{\boldsymbol{b}} \in \mathcal{B}$ and $\mathcal{B}$ is convex, L-smoothness of F on $\mathcal{B}$ gives the following quadratic upper bound

$$F(\hat{\boldsymbol{b}}) \leq F(\bar{\boldsymbol{b}}) + \left\langle \nabla F(\bar{\boldsymbol{b}}), \hat{\boldsymbol{b}} - \bar{\boldsymbol{b}} \right\rangle + \frac{L}{2} \|\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}\|_2^2. \quad (29)$$

Let $\boldsymbol{\delta} := \hat{\boldsymbol{b}} - \bar{\boldsymbol{b}} = \boldsymbol{\xi} - \boldsymbol{r}$. Since $\boldsymbol{1}^\top \boldsymbol{\xi} = \boldsymbol{1}^\top \boldsymbol{r} = R_{\text{rem}}$, we have $\boldsymbol{1}^\top \boldsymbol{\delta} = 0$. Let $(\lambda, \boldsymbol{\mu})$ be KKT multipliers for the constraints $\boldsymbol{1}^\top \boldsymbol{b} \leq B$ and $\boldsymbol{b} \geq \boldsymbol{0}$, so

$$\nabla F(\bar{\boldsymbol{b}}) + \lambda \boldsymbol{1} - \boldsymbol{\mu} = \boldsymbol{0}, \quad \lambda \geq 0, \quad \boldsymbol{\mu} \geq \boldsymbol{0}, \quad \mu_i \bar{b}_i = 0 \quad \forall i.$$

If $r_i > 0$, then $\bar{b}_i > 0$, so complementary slackness gives $\mu_i = 0$. If $r_i = 0$, then $\delta_i = \xi_i - r_i = 0$ because largest remainder rounding only rounds up coordinates with positive fractional part. Therefore

$$\left\langle \nabla F(\bar{\boldsymbol{b}}), \boldsymbol{\delta} \right\rangle = \sum_{i=1}^m (-\lambda + \mu_i) \delta_i$$
$$= -\lambda \sum_{i=1}^m \delta_i = -\lambda \boldsymbol{1}^\top \boldsymbol{\delta} = 0.$$

Substituting into (29) and applying Proposition 4 yields

$$F(\hat{\boldsymbol{b}}) \leq F(\bar{\boldsymbol{b}}) + \frac{L}{2} \|\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}\|_2^2 \leq F(\bar{\boldsymbol{b}}) + \frac{L}{2} \sum_{i=1}^m r_i(1 - r_i),$$

which proves (27). The simplified bound follows from $r_i(1 - r_i) \leq r_i$ and $r_i(1 - r_i) \leq 1/4$ for each i. ∎

Remark. The bound in Theorem 2 requires KKT stationarity of the relaxed solution, which is satisfied by interior point solutions. Frank–Wolfe only guarantees a small FW gap, so the bound applies approximately. In practice, the bound depending on the global Lipschitz constant can be loose; some future work would be to tighten this bound using local Lipschitz constant or specific structural properties of a given problem's sensing matrix such as sparsity.

V. Numerical Experiments

We evaluate the proposed Frank–Wolfe algorithm (Algorithm 1) and compare it against an interior point method (Ipopt) with analytic gradient and L-BFGS Hessian approximation. Both methods use the gradient computation from Lemma 1. All experiments were conducted on an AMD Ryzen 5 7600X3D (6 cores, 4.1 GHz) with 48 GB RAM using Julia 1.12.

A. Comparison of Solver Computation Time

a) Solver Configuration: The proposed method uses the FW algorithm with the short step rule (22), which computes the step size as $\gamma_t = \min\{g_t/(2LB^2), 1\}$ using the analytic Lipschitz constant $L = (\ln 4)^2 \|\boldsymbol{C}_x\|_2 (2m+1)$. The solver runs for at most 500 iterations with a duality gap tolerance of 10^{-6} and a 600 s wall-clock time limit. Each iteration requires one Cholesky factorization of the $d \times d$ information matrix $\boldsymbol{M}(\boldsymbol{b})$ via LAPACK, and the information matrix is assembled using a BLAS symmetric rank-k update.

The interior point baseline uses Ipopt [25] via JuMP [26] with the analytic gradient from Lemma 1 and L-BFGS Hessian approximation, initialized at $\boldsymbol{b} = (B/m)\boldsymbol{1}$ with a 600 s time limit. Each Ipopt iteration requires the same $O(d^3)$ Cholesky factorization as FW. After the continuous solve, both methods apply the same largest remainder rounding procedure from Algorithm 2 to obtain integer bit allocations.

b) Test Cases: We use 11 power grid test cases from the PGLib-OPF benchmark library [29]. The sensing matrix $\boldsymbol{H}$ is the grounded bus susceptance matrix from DC power flow with the slack bus removed to ensure positive definiteness. State dimensions range from $d = 13$ (case14) to $d = 299$ (case300), with $m = d$ measurements corresponding to power injections at all non slack buses. For each test case, we generate 30 independent problem instances by randomizing the channel precision constants $\kappa_i \sim \text{Uniform}(0.8, 1.2)$, $i = 1, \ldots, m$. The bit budget is set to $B = 2m$ (2 bits per sensor). The prior covariance is $\boldsymbol{C}_x = \boldsymbol{I}$.

Table I summarizes solve times across the 30 randomized instances for each test case. Both methods are bottlenecked by one LAPACK Cholesky factorization of the information matrix $\boldsymbol{M}(\boldsymbol{b})$ per iteration, with the gradient extracted via $[\boldsymbol{H}\boldsymbol{C}_\varepsilon^2\boldsymbol{H}^\top]_{ii}$ without forming the full $m \times m$ product. The FW LMO adds only $O(m)$ work per iteration via (20); the interior point KKT solve is similarly $O(m)$ with L-BFGS. The speed difference is determined by iteration count; the interior point method converges in 20–50 iterations, while FW often used most of the max 500 iterations. Across all instances where both solvers converged, FW and Ipopt+∇ returned similar objective values up to the solver tolerances.
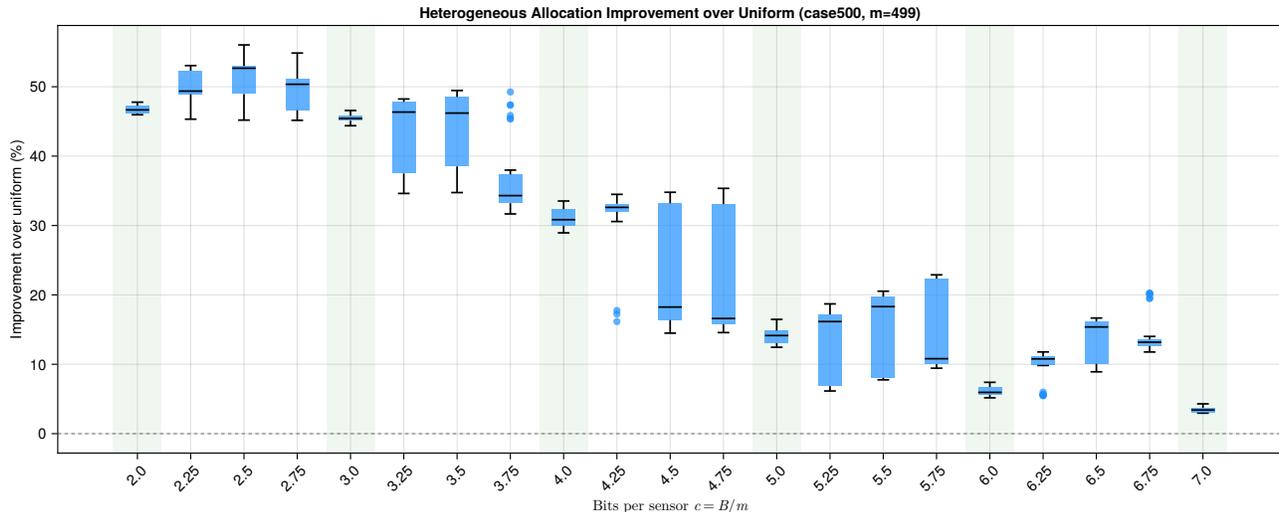
Fig. 3. Percentage improvement of the optimized heterogeneous bit allocation (8) over uniform allocation $\boldsymbol{b} = \boldsymbol{1} \cdot \lfloor B/m \rfloor$ on the `case500` test system ($m = 499$ sensors). Each boxplot shows 30 randomized instances at the given budget level $c = B/m$. Shaded columns denote integer values of c.

TABLE I

SOLVE TIME COMPARISON: STOCK IPOPT, FW, AND IPOPT + ANALYTIC
GRADIENT. MEAN ± STD. DEV. OVER 30 TRIALS, 10 MIN. LIMIT.

Case	m	Ipopt (s)	FW (s)	Ipopt+∇ (s)
case14	13	5.03 ± 1.02	0.098 ± 0.0023	0.033 ± 0.14
case30	29	2.73 ± 0.16	0.24 ± 0.0093	0.014 ± 0.0010
case57	56	7.52 ± 0.31	0.89 ± 0.030	0.022 ± 0.0020
case73	72	12.1 ± 0.97	1.76 ± 0.090	0.026 ± 0.0040
case118	117	121.2 ± 12.6	7.64 ± 0.22	0.052 ± 0.0040
case162	161	252.9 ± 26.1	27.44 ± 1.1	0.070 ± 0.0060
case179	178	442.7 ± 80.3	41.05 ± 1.6	0.16 ± 0.032
case197	196	> 600 (timeout)	50.67 ± 1.5	0.16 ± 0.020
case200	199	> 600 (timeout)	52.59 ± 2.7	0.12 ± 0.028
case240	239	> 600 (timeout)	77.82 ± 2.4	0.11 ± 0.056
case300	299	> 600 (timeout)	207.46 ± 8.6	0.43 ± 0.062

TABLE II

MEDIAN EXPERIMENTAL ROUNDING QUALITY VS THEORETICAL BOUND
OVER 30 RANDOM INSTANCES AND $B = 2m$ TOTAL BIT BUDGET.

Case	m	$F(\hat{\boldsymbol{b}}) - F(\bar{\boldsymbol{b}})$	Rounding bound (27)	Median ratio
case14	13	1.92×10^{-2}	7.26×10^{1}	2.62×10^{-4}
case30	29	2.48×10^{-2}	2.52×10^{2}	9.98×10^{-5}
case57	56	4.13×10^{-2}	1.00×10^{3}	4.06×10^{-5}
case200	199	5.76×10^{-2}	1.34×10^{4}	4.28×10^{-6}
case240	239	1.80×10^{-2}	1.57×10^{4}	1.15×10^{-6}
case300	299	8.38×10^{-2}	2.73×10^{4}	3.05×10^{-6}

Table I therefore isolates runtime differences rather than solution quality differences.

Table II reports the median rounding gap $F(\hat{\boldsymbol{b}}) - F(\bar{\boldsymbol{b}})$ and the theoretical bound from Theorem 2 across six IEEE test cases, each evaluated over 30 random instances with $\kappa_i \sim \text{Uniform}(0.8, 1.2)$. In particular, the ratio of the actual gap to the bound ranges from 10^{-4} (case14) to 10^{-6} (case240, case300), indicating that the bound is highly conservative in practice. This conservatism may stem from L being a global worst-case Lipschitz constant, whereas a *local* Lipschitz constant could give a tighter bound.

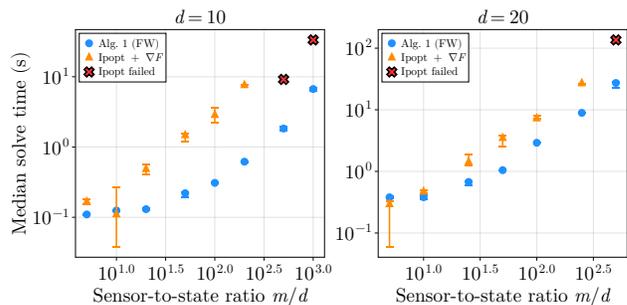In Figure 4, we analyzed a "sensor rich" ($m \gg d$) regime

Fig. 4. Median solve time vs. sensor-to-state ratio m/d for $d \in \{10, 20\}$ with $B = 2d$. Error bars show the IQR over 30 instances. Frank–Wolfe scales linearly in m (for $m \gg d$) with $O(d^3 + d^2 m)$ per-iteration cost. Ipopt fails to converge entirely beyond $m/d \approx 200$ ($d{=}10$) and 100 ($d{=}20$).

in which FW outperforms Ipopt with gradient acceleration in computation time. We generate random Gaussian instances with $\boldsymbol{H} \in \mathbb{R}^{m \times d}$, $\kappa_i \sim \mathcal{U}(0.8, 1.2)$, and bit budget $B = 2d$. For each state dimension $d \in \{10, 20\}$, we sweep the sensor-to-state ratio m/d from 5 to 1000 and solve 30 independent instances per configuration using both Algorithm 1 and Ipopt with analytic gradients and L-BFGS. Median solve times and ranges are reported; red × markers indicate configurations where Ipopt failed to converge on all 30 instances.

B. Comparison of Problem Formulations

We evaluate the benefit of solving (8) relative to the uniform baseline $\boldsymbol{b} = \boldsymbol{1} \cdot \lfloor B/m \rfloor$ on the `case500` power system test case ($m = d = 499$) across per-sensor budgets $c = B/m \in [2, 7]$. For each budget level, 30 instances are generated with quantization gains $\kappa_i \sim \text{Uniform}(0.8, 1.2)$, and the continuous relaxation is solved with Ipopt+∇ method followed by largest remainder rounding.

As shown in Figure 3, heterogeneous allocation yields median improvements of approximately 47% at $c = 2$, rising to 53% at $c = 2.5$ before settling near 50% and 45%

at $c = 2.75$ and $c = 3$, respectively, then declining to 31% at $c = 4$ and 14% at $c = 5$, and falling to 3.4% at $c = 7$ as abundant access to bandwidth levels the playing field. The gains at integer c remain large (e.g., 45% at $c = 3$, 31% at $c = 4$), confirming that the dominant source of improvement is the optimizer's ability to concentrate precision on high information sensors rather than an artifact of baseline rounding. These results demonstrate that heterogeneous allocation is most valuable in the bandwidth-constrained regime that motivates this work.

VI. CONCLUSIONS

We proposed an optimal bit allocation method for state estimation with variable precision measurements. The method reduces the MSE of a state estimator by controlling how a limited number of bits are allocated to sensors.

a) Discussion: Two algorithms were proposed: a first-order and second-order method to solve the corresponding nonconvex optimization problem. The algorithms demonstrated the ability to heterogeneously allocate a limited bit budget across quantized measurements, significantly reducing the A-optimal design criterion for the state estimator. We derived a closed-form gradient formula that reduces computing the gradient to a single Cholesky factorization. We presented two methods that exploit this gradient: a Frank–Wolfe method with a closed-form LMO and guaranteed $O(1/\sqrt{T})$ convergence rate guarantee in terms of the FW gap, and an interior point method with L-BFGS Hessian approximation. Numerical experiments on IEEE power grid test cases show that the interior point method converges in fewer iterations, as expected for a second-order method. Additionally, the Frank–Wolfe method, being a first order method compared to a second order interior point method, is more memory efficient, allowing it to scale to large problems where storing the L-BFGS Hessian itself becomes infeasible.

b) Future work: This work opens a wide array of future directions. This includes adding stochastic trace estimation for scaling to larger problem sizes, such as the classical Hutchinson estimator or more modern approaches [30]. This paper has thus far considered only an offline state estimation problem. Future work should extend these methods to dynamic bit allocation in time-varying systems. Another promising direction is the consideration of a broader class of objectives that measure communication cost and performance criteria complementary to the A-optimal design criterion. Future work will consider general classes of convex communication cost functionals, and applying decision-focused learning (DFL) principles to optimize over the dynamic ranges of the channels.

REFERENCES

[1] R. Gray and D. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[2] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A Survey of Quantization Methods for Efficient Neural Network Inference," in *Low-Power Computer Vision*. Chapman and Hall/CRC, 2022, pp. 291–326.

[3] Y. Plan and R. Vershynin, "The Generalized LASSO with Non-Linear Observations," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1528–1537, 2016.

[4] C. Thrampoulidis and A. S. Rawat, "The Generalized LASSO for Sub-Gaussian Measurements with Dithered Quantization," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2487–2500, 2020.

[5] S. Joshi and S. Boyd, "Sensor Selection via Convex Optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, 2009.

[6] J. Wang, W. Xie, and I. O. Ryzhov, "Algorithms for Budget-Constrained D-Optimal Design," *Math. Oper. Res.*, 2025, early access.

[7] A. Krause, J. Leskovec, C. Guestrin, J. VanBriesen, and C. Faloutsos, "Efficient Sensor Placement Optimization for Securing Large Water Distribution Networks," *J. Water Resour. Plan. Manag.*, vol. 134, no. 6, pp. 516–526, 2008.

[8] L. F. O. Chamon and A. Ribeiro, "Approximate Supermodularity Bounds for Experimental Design," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5409–5418.

[9] V. Madan, M. Singh, U. Tantipongpipat, and W. Xie, "Combinatorial Algorithms for Optimal Design," in *Proc. 32nd COLT*, vol. 99, 2019, pp. 2210–2258.

[10] A. Nikolov, M. Singh, and U. T. Tantipongpipat, "Proportional Volume Sampling and Approximation Algorithms for A-Optimal Design," in *Proc. 30th ACM-SIAM SODA*, 2019, pp. 1369–1386.

[11] M. Frank and P. Wolfe, "An Algorithm for Quadratic Programming," *Nav. Res. Logist. Q.*, vol. 3, no. 1–2, pp. 95–110, 1956.

[12] S. D. Ahipasaoglu, "A First-Order Algorithm for the A-Optimal Experimental Design Problem: A Mathematical Programming Approach," *Stat. Comput.*, vol. 25, no. 6, pp. 1113–1127, Nov. 2015.

[13] F. de la Hucha Arce, P. Patrinos, M. Verhelst, and A. Bertrand, "On the Convexity of Bit Depth Allocation for Linear MMSE Estimation in Wireless Sensor Networks," *IEEE Signal Process. Lett.*, vol. 27, pp. 291–295, 2020.

[14] F. de la Hucha Arce, F. Rosas, M. Moonen, M. Verhelst, and A. Bertrand, "Generalized Signal Utility for LMMSE Signal Estimation With Application to Greedy Quantization in Wireless Sensor Networks," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1202–1206, 2016.

[15] V. V. Fedorov, *Theory of Optimal Experiments*. Academic Press, Inc., 1972.

[16] S. Sun, J. Lin, L. Xie, and W. Xiao, "Quantized Kalman Filtering," in *Proc. 22nd IEEE ISIC*, 2007, pp. 7–12.

[17] E. J. Msechu, S. I. Roumeliotis, A. Ribeiro, and G. B. Giannakis, "Decentralized Quantized Kalman Filtering with Scalable Communication Cost," *IEEE Trans. Signal Process.*, vol. 56, no. 8, pp. 3727–3741, 2008.

[18] V. Kekatos, G. B. Giannakis, and B. Wollenberg, "Optimal Placement of Phasor Measurement Units via Convex Relaxation," *IEEE Trans. Power Syst.*, vol. 27, no. 3, pp. 1521–1530, 2012.

[19] S. Talkington, A. Rangarajan, P. A. Alcântara, L. Roald, D. K. Molzahn, and D. R. Fuhrmann, "Error Bounds for Radial Network Topology Learning from Quantized Measurements," *IEEE Trans. Power Syst.*, pp. 1–4, 2026.

[20] G. Cheng, Y. Lin, A. Abur, A. Gómez-Expósito, and W. Wu, "A Survey of Power System State Estimation using Multiple Data Sources: PMUs, SCADA, AMI, and Beyond," *IEEE Trans. Smart Grid*, vol. 15, no. 1, pp. 1129–1151, 2024.

[21] R. Gray and T. Stockham, "Dithered Quantizers," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 805–812, 1993.

[22] M. Jaggi, "Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization," in *Proc. 30th ICML*, vol. 28, no. 1, 2013, pp. 427–435.

[23] S. Pokutta, "The Frank-Wolfe Algorithm: A Short Introduction," *Jahresber. Dtsch. Math.-Ver.*, vol. 126, no. 1, pp. 3–35, Mar. 2024.

[24] S. Lacoste-Julien, "Convergence Rate of Frank-Wolfe for Non-Convex Objectives," *arXiv:1607.00345 [math.OC]*, 2016.

[25] A. Wächter and L. T. Biegler, "On the Implementation of an Interior-Point Filter Line-Search Algorithm for Large-Scale Nonlinear Programming," *Math. Program.*, vol. 106, no. 1, pp. 25–57, Mar. 2006.

[26] I. Dunning, J. Huchette, and M. Lubin, "JuMP: A Modeling Language for Mathematical Optimization," *SIAM Rev.*, vol. 59, no. 2, pp. 295–320, 2017.

[27] M. L. Balinski and H. P. Young, "The Quota Method of Apportionment," *Amer. Math. Monthly*, vol. 82, no. 7, pp. 701–730, 1975.

[28] F. Pukelsheim and S. Rieder, "Efficient Rounding of Approximate Designs," *Biometrika*, vol. 79, no. 4, pp. 763–770, 1992.

[29] S. Babaeinejadsarookolaee *et al.*, "The Power Grid Library for Benchmarking AC Optimal Power Flow Algorithms," *arXiv:1908.02788 [math.OC]*, 2021.

[30] R. A. Meyer, C. Musco, C. Musco, and D. P. Woodruff, "Hutch++: Optimal Stochastic Trace Estimation," in *Proc. SIAM SOSA*, 2021, pp. 142–155.