Power Diagrams for Embedding-Based Ad Auctions: Mechanism Design in Continuous Intent Space

June Kim
june@june.kim

February 2026

Abstract

The shift from keyword-based to embedding-based ad targeting in large language model platforms creates a fundamental challenge: advertisers must bid on *regions* of continuous, high-dimensional space rather than discrete keywords. We propose a geometric framework based on *power diagrams* (additively weighted Voronoi tessellations) where each advertiser's territory in embedding space is determined by their bid-weighted proximity. Our main theoretical contribution is establishing that, for isotropic Gaussian value functions over a continuous impression space, the welfare-maximizing allocation *is* a power diagram, and the unique incentive-compatible, individually rational, welfare-maximizing mechanism computes payments via Voronoi cell integration—reducing the mechanism design problem to a computational geometry problem (Theorem 3.3). We characterize the geometry of strategic behavior: under exact VCG, center misreporting never helps (Theorem 3.6), but under approximate payment rules, the gradient of value received points toward high-density regions of the impression distribution (Theorem 3.5), quantifying the vulnerability of practical mechanisms to spatial manipulation. We prove that best-response bid dynamics converge to the unique Nash equilibrium in N rounds (Theorem 3.8), and show that random projections to $O(\log N/\varepsilon^2)$ dimensions preserve the allocation (Theorem 5.3), making high-dimensional deployment tractable. Experimentally, we compare power-diagram-native auctions against discretized baselines (GSP and VCG on IAB-style category grids) across $d \in \{2, 5, 10, 20\}$. At $d = 2$, all mechanisms perform similarly, but the welfare gap widens dramatically with dimensionality: Power-VCG achieves 37% higher social welfare than discretized baselines at $d = 10$ and 32% at $d = 20$, while maintaining exactly zero IC regret (versus regret exceeding 1.0 for discretized VCG at $d \geq 10$). Winner determination via kd-tree indexing runs in $\sim 0.15\,\mu$s per impression, independent of the number of advertisers, meeting real-time latency requirements.

1 Introduction

Online advertising has been built on discrete auction mechanisms. In Google's search advertising, the fundamental biddable unit is a *keyword*: advertisers bid on "running shoes" or "mesothelioma lawyer," and a second-price-like auction clears efficiently for each query. This discrete structure enables clean market mechanics—winner determination, truthful bidding incentives, and tractable budget forecasting.

The emergence of large language model (LLM) platforms as advertising venues disrupts this foundation. OpenAI announced in January 2026 that it would begin testing advertisements in ChatGPT, describing their approach as "intent-based monetization" across 800 million monthly users [OpenAI, 2026]. Similar efforts are underway at Perplexity and other conversational AI platforms.

In these systems, user interactions are not represented as keywords but as points in continuous, high-dimensional *embedding spaces*. A user's conversation traces a trajectory through this space, and an advertising "impression" corresponds to a point encoding semantic content, user intent, temporal context, and other dimensions simultaneously. Advertisers do not want individual keywords—they want *regions* of this space.

This paper introduces a geometric framework for ad auctions in continuous embedding spaces. Our key insight is that when advertiser value functions are modeled as functions of distance in embedding space, the natural allocation mechanism is a *power diagram*—an additively weighted Voronoi tessellation where bids determine territory boundaries.

Contributions.

1. We formalize the embedding-based ad auction problem, defining impression spaces, advertiser value functions, and allocation mechanisms in continuous space (Section 2).

2. For the isotropic Gaussian case, we prove the welfare-maximizing allocation is a power diagram and that the mechanism with VCG payments computed via Voronoi cell integration is the *unique* IC, IR, welfare-maximizing mechanism—reducing mechanism design to computational geometry (Section 3, Theorem 3.3).

3. We characterize the geometry of strategic behavior: exact VCG is fully IC for center reports (Theorem 3.6), but under approximate payment rules, the value-received gradient (Theorem 3.5) quantifies vulnerability to spatial manipulation—advertisers are incentivized to shift toward high-density impression regions.

4. We prove that best-response bid dynamics converge to the unique Nash equilibrium in N rounds (Theorem 3.8), and contrast this with cycling under GSP (Theorem 3.9).

5. We extend the framework to anisotropic preferences with quadric boundaries (Section 4) and mixture-of-Gaussian preferences with virtual advertiser decomposition (Section 5).

6. We show that random projection to $O(\log N/\varepsilon^2)$ dimensions preserves the power diagram allocation (Theorem 5.3), making high-dimensional deployment tractable.

7. We provide $O(\log N)$ winner determination and payment computation via spatial indexing, with a full auction algorithm (Section 6).

8. We give budget prediction bounds via Monte Carlo estimation (Section 7).

9. We experimentally validate the framework against discretized baselines, showing consistent improvements in welfare and revenue (Section 8).

Related work. Our work bridges several lines of research. We describe each precisely to clarify our contribution.

Mechanism design for LLMs. Dütting et al. [2024b] model LLM interactions as sequential token-level decisions and design mechanisms for selecting which tokens to monetize; their allocation unit is a token or token sequence. Hajiaghayi et al. [2024] propose RAG-based ad auctions where retrieved documents carry ads and the mechanism selects which documents to surface; the biddable unit is a document slot in the retrieval pipeline. Xu et al. [2026] model genre-based bidding for LLM-generated responses, discretizing the ad space into a fixed set of content genres. Banchio et al. [2025] study ads in conversational settings with a focus on user experience externalities across turns. Our

model operates at the impression level: each conversation maps to a point in continuous embedding space, and the mechanism allocates that point to an advertiser via geometric partitioning. The key difference is that token-level and slot-level mechanisms cannot express spatial competition between advertisers over regions of semantic space, which is the central object of our framework.

Optimal transport and mechanism design. Daskalakis et al. [2013] establish that optimal multi-item auction design with continuous buyer types can be cast as an optimal transport problem, yielding characterizations of revenue-optimal mechanisms. Their framework applies to a single buyer with multidimensional types. Our setting differs: we have N buyers (advertisers) competing over a continuum of items (impressions), and we seek welfare-maximizing rather than revenue-optimal mechanisms. The connection is conceptual—both exploit continuous-space structure—but the technical tools differ: we use power diagrams rather than optimal transport duality.

Computational geometry of auctions. Joswig et al. [2024] characterize the polyhedral structure of the set of all truthful auction mechanisms for finite type spaces, showing it forms a polytope whose vertices correspond to extremal mechanisms. Our work is complementary: we show that for a specific class of continuous-type auctions (Gaussian values in embedding space), the welfare-maximizing mechanism has the geometric structure of a power diagram [Aurenhammer, 1987], enabling efficient computation. Power diagrams have not previously been connected to ad auction design.

Neural auction design. Dütting et al. [2024a] learn approximately optimal (revenue-maximizing) auction rules via deep networks, training on sampled bidder valuations. Liu et al. [2021] deploy learned auction mechanisms at Alibaba, jointly optimizing allocation and payment rules. These approaches learn the auction mechanism end-to-end but treat the targeting problem (which ad is relevant to which user) as exogenous. Our framework unifies targeting and auction design: the power diagram simultaneously determines relevance (geometric proximity) and allocation (cell membership), with provable incentive guarantees that learned mechanisms lack.

Industry practice. Current contextual advertising discretizes embedding spaces into taxonomies (IAB's ∼700 content categories) and runs standard real-time bidding (RTB) auctions independently per cell [Seedtag, 2025, GumGum/Verity, 2024]. This discretization discards the continuous structure of embeddings: two impressions on opposite sides of a category boundary may be nearly identical in embedding space yet participate in entirely different auctions. Our power diagram approach operates directly on the continuous space, eliminating boundary artifacts and capturing the full geometric structure of advertiser competition.

2 Model

Definition 2.1 (Impression Space)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a compact convex set representing the embedding space. Points $\mathbf{x} \in \mathcal{X}$ represent impressions—user interactions projected into embedding space. The impression distribution μ is a probability measure on $\mathcal{X}$ representing the density of user traffic.*

Definition 2.2 (Advertiser)**.** *An advertiser $i \in [N] = \{1, \ldots, N\}$ is characterized by:*

- *A center $\mathbf{c}_i \in \mathcal{X}$ (their ideal customer embedding),*
- *A bid scalar $b_i \in \mathbb{R}_{>0}$ (willingness to pay),*
- *A value function $v_i : \mathcal{X} \to \mathbb{R}_{\geq 0}$ parameterized by the above.*

Definition 2.3 (Allocation)**.** *An allocation is a measurable partition $\{V_0, V_1, \ldots, V_N\}$ of $\mathcal{X}$, where V_i is the set of impressions allocated to advertiser i, and V_0 is the unallocated set (reserved, e.g., for policy exclusions).*

The *social welfare* of an allocation is:

$$W(\{V_i\}) = \sum_{i=1}^{N} \int_{V_i} v_i(\mathbf{x})\, d\mu(\mathbf{x}). \tag{1}$$

The welfare-maximizing allocation assigns each impression to the advertiser who values it most:

$$i^*(\mathbf{x}) = \arg\max_{i \in [N]} v_i(\mathbf{x}). \tag{2}$$

Definition 2.4 (Payment Rule)**.** *A payment rule $p = (p_1, \ldots, p_N)$ specifies each advertiser's total payment. A mechanism $(\{V_i\}, p)$ is* incentive compatible *(IC) if truthful reporting of type parameters (bid b_i, center $\mathbf{c}_i$, and covariance Σ_i as applicable) is a dominant strategy for each advertiser.*

We consider three increasingly expressive parameterizations of v_i:

Isotropic Gaussian. $v_i(\mathbf{x}) = b_i \cdot \exp\!\Big(-\frac{\|\mathbf{x}-\mathbf{c}_i\|^2}{\sigma_i^2}\Big)$, where $\sigma_i > 0$ is a reach parameter. When $\sigma_i = \sigma$ for all i (shared reach), the allocation problem has power diagram structure.

Anisotropic Gaussian. $v_i(\mathbf{x}) = b_i \cdot \exp\!\big(-(\mathbf{x}-\mathbf{c}_i)^\top \Sigma_i^{-1} (\mathbf{x}-\mathbf{c}_i)\big)$, where $\Sigma_i \succ 0$ is a covariance matrix encoding directional preferences.

Mixture of Gaussians. $v_i(\mathbf{x}) = b_i \cdot \sum_{k=1}^{K_i} w_{ik} \cdot \exp\!\big(-(\mathbf{x}-\mathbf{c}_{ik})^\top \Sigma_{ik}^{-1} (\mathbf{x}-\mathbf{c}_{ik})\big)$, representing multimodal preferences.

Notation summary. Table 1 collects the notation used throughout the paper.

3 Isotropic Case: Power Diagrams

We first analyze the case where all advertisers share a common reach parameter $\sigma > 0$, so $v_i(\mathbf{x}) = b_i \cdot \exp(-\|\mathbf{x}-\mathbf{c}_i\|^2/\sigma^2)$.

Lemma 3.1 (Log-value characterization)**.** *The welfare-maximizing allocation assigns impression $\mathbf{x}$ to $i^*(\mathbf{x}) = \arg\max_i v_i(\mathbf{x})$. Since $\exp(\cdot)$ is monotone and σ is shared:*

$$i^*(\mathbf{x}) = \arg\max_{i \in [N]} \left[\log b_i - \frac{\|\mathbf{x}-\mathbf{c}_i\|^2}{\sigma^2} \right]. \tag{3}$$

This is precisely the definition of a *power diagram* (additively weighted Voronoi diagram) with sites $\mathbf{c}_i$ and weights $w_i = \sigma^2 \log b_i$.

Definition 3.2 (Power Diagram)**.** *Given sites $\mathbf{c}_1, \ldots, \mathbf{c}_N \in \mathbb{R}^d$ and weights $w_1, \ldots, w_N \in \mathbb{R}$, the power diagram is the partition $\{V_1, \ldots, V_N\}$ where:*

$$V_i = \Big\{ \mathbf{x} \in \mathcal{X} : \|\mathbf{x}-\mathbf{c}_i\|^2 - w_i \le \|\mathbf{x}-\mathbf{c}_j\|^2 - w_j \;\; \forall j \neq i \Big\}. \tag{4}$$

The boundary between cells V_i and V_j is the hyperplane:

$$H_{ij} = \Big\{ \mathbf{x} : \|\mathbf{x}-\mathbf{c}_i\|^2 - w_i = \|\mathbf{x}-\mathbf{c}_j\|^2 - w_j \Big\} = \Big\{ \mathbf{x} : 2(\mathbf{c}_j - \mathbf{c}_i)^\top \mathbf{x} = \|\mathbf{c}_j\|^2 - \|\mathbf{c}_i\|^2 - w_j + w_i \Big\}. \tag{5}$$

Table 1: Summary of notation.

Symbol	Meaning
$\mathcal{X} \subseteq \mathbb{R}^d$	Impression (embedding) space
$\mathbf{x} \in \mathcal{X}$	A single impression
μ	Impression distribution (probability measure on $\mathcal{X}$)
N	Number of advertisers
$\mathbf{c}_i \in \mathcal{X}$	Center of advertiser i (ideal customer embedding)
$b_i \in \mathbb{R}_{>0}$	Bid scalar of advertiser i
σ, σ_i	Reach parameter(s) (isotropic Gaussian bandwidth)
$\Sigma_i \succ 0$	Covariance matrix of advertiser i (anisotropic case)
$v_i(\mathbf{x})$	Value of impression $\mathbf{x}$ to advertiser i
$w_i = \sigma^2 \log b_i$	Power diagram weight for advertiser i
V_i	Power diagram cell (territory) of advertiser i
V_j^{-i}	Cell of advertiser j in the diagram computed without i
$W(\{V_i\})$	Social welfare of allocation $\{V_i\}$
p_i	Payment of advertiser i
u_i	Utility of advertiser i: value received minus payment
H_{ij}	Hyperplane boundary between cells V_i and V_j
K_i	Number of mixture components for advertiser i
w_{ik}	Mixture weight for component k of advertiser i
B_i	Budget of advertiser i

Theorem 3.3 (Power Diagram Mechanism: Uniqueness, IC, and IR)**.** *Consider N advertisers with isotropic Gaussian value functions sharing a common reach parameter $\sigma > 0$, over a continuous impression space $(\mathcal{X}, \mu)$. Define the* power diagram mechanism *$\mathcal{M} = (\{V_i\}, p)$ as follows:*

1. ***Allocation.*** *The allocation is the power diagram with sites $\mathbf{c}_i$ and weights $w_i = \sigma^2 \log b_i$:*

$$V_i = \left\{\mathbf{x} \in \mathcal{X} : \|\mathbf{x} - \mathbf{c}_i\|^2 - w_i \leq \|\mathbf{x} - \mathbf{c}_j\|^2 - w_j \ \ \forall j \neq i\right\}. \tag{6}$$

2. ***Payment.*** *Each advertiser i pays the VCG externality, computed via Voronoi cell integration:*

$$p_i = \sum_{j \neq i} \int_{V_j^{-i}} v_j(\mathbf{x})\, d\mu(\mathbf{x}) - \sum_{j \neq i} \int_{V_j} v_j(\mathbf{x})\, d\mu(\mathbf{x}), \tag{7}$$

where $\{V_j^{-i}\}_{j \neq i}$ is the power diagram computed on the $N-1$ advertisers excluding i.

Then $\mathcal{M}$ satisfies:

(a) ***Welfare maximization.*** *$\{V_i\}$ maximizes social welfare $W(\{V_i\}) = \sum_i \int_{V_i} v_i(\mathbf{x})\, d\mu(\mathbf{x})$ over all measurable partitions of $\mathcal{X}$.*

(b) ***Incentive compatibility.*** *Truthful bid reporting is a dominant strategy: for each i and any misreport $b_i' \neq b_i$, $u_i(b_i) \geq u_i(b_i')$.*

(c) ***Individual rationality.*** *Every advertiser has non-negative utility: $u_i \geq 0$ for all i.*

(d) ***Uniqueness.*** *$\mathcal{M}$ is the unique mechanism (up to tie-breaking on measure-zero sets) that simultaneously satisfies welfare maximization, IC, and IR, among mechanisms where payments do not depend on advertiser i's own report beyond the allocation.*

Proof. We establish each claim. The core contribution is showing that the continuous-space mechanism design problem reduces to a computational geometry problem: welfare maximization becomes power diagram construction, and payment computation becomes Voronoi cell integration.

Part (a): Welfare maximization. The welfare-maximizing allocation assigns each impression pointwise to the highest-value advertiser: $i^*(\mathbf{x}) = \arg\max_i v_i(\mathbf{x})$. For isotropic Gaussians with shared σ, we have $v_i(\mathbf{x}) = b_i \exp(-\|\mathbf{x} - \mathbf{c}_i\|^2 / \sigma^2)$. Since $\exp(\cdot)$ is strictly monotone and σ is shared:

$$i^*(\mathbf{x}) = \arg\max_i \left[\log b_i - \|\mathbf{x} - \mathbf{c}_i\|^2 / \sigma^2\right] = \arg\min_i \left[\|\mathbf{x} - \mathbf{c}_i\|^2 - \sigma^2 \log b_i\right]. \tag{8}$$

This is precisely the definition of the power diagram with sites $\mathbf{c}_i$ and weights $w_i = \sigma^2 \log b_i$ [Aurenhammer, 1987]. The reduction from welfare maximization to power diagram construction is exact: no approximation or discretization is involved. The partition $\{V_i\}$ defined by (6) implements the pointwise maximum, hence maximizes welfare.

Part (b): Incentive compatibility. We adapt the VCG argument [Vickrey, 1961, Clarke, 1971, Groves, 1973] to the continuous-impression setting. Write advertiser i's utility under truthful reporting:

$$\begin{aligned} u_i(b_i) &= \int_{V_i} v_i(\mathbf{x})\, d\mu(\mathbf{x}) - p_i \\ &= \int_{V_i} v_i(\mathbf{x})\, d\mu(\mathbf{x}) - \left[\sum_{j \neq i} \int_{V_j^{-i}} v_j(\mathbf{x})\, d\mu(\mathbf{x}) - \sum_{j \neq i} \int_{V_j} v_j(\mathbf{x})\, d\mu(\mathbf{x})\right] \\ &= \underbrace{\sum_{j=1}^{N} \int_{V_j} v_j(\mathbf{x})\, d\mu(\mathbf{x})}_{=W(\{V_k\})} - \underbrace{\sum_{j \neq i} \int_{V_j^{-i}} v_j(\mathbf{x})\, d\mu(\mathbf{x})}_{h_i \text{ (independent of } i\text{'s report)}}. \end{aligned} \tag{9}$$

The term $h_i = \sum_{j \neq i} \int_{V_j^{-i}} v_j(\mathbf{x})\, d\mu(\mathbf{x})$ depends only on the other advertisers' types and is independent of i's reported bid. Hence i maximizes u_i by maximizing $W(\{V_k\})$. Since the allocation rule already maximizes W when i reports truthfully, any misreport $b_i' \neq b_i$ can only decrease W (or leave it unchanged), so $u_i(b_i) \geq u_i(b_i')$.

Part (c): Individual rationality. We must show $u_i \geq 0$. By (9), $u_i = W(\{V_k\}) - h_i$, where $W(\{V_k\})$ is the maximum welfare with all N advertisers and h_i is the maximum welfare achievable by the $N-1$ advertisers excluding i. Adding advertiser i to the market can only increase (or maintain) welfare, since one could always allocate nothing to i and recover the $(N-1)$-advertiser optimum. Formally, $W(\{V_k\}) \geq \sum_{j \neq i} \int_{V_j^{-i}} v_j(\mathbf{x})\, d\mu(\mathbf{x}) = h_i$ because $\{V_j^{-i}\}_{j \neq i} \cup \{V_0^{-i}\}$ is a feasible (but not necessarily optimal) partition when i is present (assigning $V_0^{-i} \cup V_i^{-i}$ as unallocated or redistributed). Hence $u_i \geq 0$.

Part (d): Uniqueness. By the Green–Laffont characterization [Green and Laffont, 1977], in quasi-linear environments where the social choice function maximizes the sum of valuations, the Groves class of payment rules is the unique class that implements the efficient allocation in dominant strategies. The VCG mechanism (with h_i depending only on others' reports) is the unique member of the Groves class that additionally satisfies IR. The only degree of freedom is tie-breaking, which affects a measure-zero set of impressions (the power diagram boundaries H_{ij}) and hence does not affect utilities or payments under any absolutely continuous μ.

Computational reduction. The significance of Parts (a)–(d) is that they reduce the mechanism design problem entirely to computational geometry: (i) welfare-maximizing allocation reduces to constructing a power diagram, which is a standard $O(N \log N)$ operation [Aurenhammer, 1987];

(ii) payment computation reduces to integrating over Voronoi cells—specifically, computing h_i requires constructing the $(N-1)$-advertiser power diagram and integrating v_j over each cell V_j^{-i}. In the isotropic case with shared σ, these integrals involve Gaussian densities over polyhedral regions, which can be computed via Monte Carlo sampling or, in low dimensions, exact polyhedral integration. □

Remark 3.4 (Heterogeneous reach)**.** *When σ_i varies across advertisers, the allocation is still welfare-maximizing via* (2)*, but boundaries are no longer hyperplanes. The cell V_i is defined by:*

$$V_i = \left\{ \mathbf{x} : \log b_i - \frac{\|\mathbf{x}-\mathbf{c}_i\|^2}{\sigma_i^2} \geq \log b_j - \frac{\|\mathbf{x}-\mathbf{c}_j\|^2}{\sigma_j^2} \ \ \forall j \right\},$$

which gives quadric (conic) boundaries. Theorem 3.3 parts (b)–(d) still hold: the VCG argument in (9) *does not depend on the geometric structure of the cells, only on the pointwise welfare-maximization property of the allocation. The computational reduction to power diagrams, however, requires shared σ.*

3.1 Incentive Compatibility for Center Reports

Theorem 3.3 establishes IC with respect to bid misreporting. Theorem 3.6 (below) shows that exact VCG is also IC for center reports. However, practical mechanisms often use approximate payments (Theorem 6.2). The following theorem characterizes the geometry of value-received sensitivity to center perturbations, which becomes exploitable under approximate payment rules.

Theorem 3.5 (Value-Received Gradient for Center Perturbations)**.** *Consider the isotropic Gaussian model with $N \geq 2$ advertisers, shared reach σ, and VCG payments. Under any payment rule where the payment to advertiser i does not fully adjust to compensate for center misreporting (including the second-score approximation of Theorem 6.2 and GSP), there exist configurations where an advertiser can strictly increase utility by misreporting their center.*

Moreover, the utility gain from an infinitesimal center misreport $\mathbf{c}_i' = \mathbf{c}_i + \varepsilon\mathbf{d}$ (with $\|\mathbf{d}\| = 1$) is:

$$\left.\frac{\partial u_i}{\partial \mathbf{c}_i}\right|_{\mathbf{c}_i'=\mathbf{c}_i} \cdot \mathbf{d} = \int_{\partial V_i} \big[v_i(\mathbf{x}) - v_{j(i,\mathbf{x})}(\mathbf{x})\big] \frac{\partial H}{\partial \mathbf{c}_i} \cdot \mathbf{d} \, d\mu_\partial(\mathbf{x}) + \int_{V_i} \frac{\partial v_i}{\partial \mathbf{c}_i}(\mathbf{x}) \cdot \mathbf{d} \, d\mu(\mathbf{x}), \tag{10}$$

where ∂V_i is the boundary of V_i, $j(i,\mathbf{x})$ is the neighboring advertiser at boundary point $\mathbf{x}$, $\frac{\partial H}{\partial \mathbf{c}_i}$ is the rate of boundary displacement, and $d\mu_\partial$ is the induced boundary measure.

Under truthful reporting, $v_i(\mathbf{x}) = v_{j(i,\mathbf{x})}(\mathbf{x})$ on ∂V_i (by the power diagram construction), so the first integral vanishes, and the gradient of value received *with respect to center perturbation is:*

$$\nabla_{\mathbf{c}_i}\left[\int_{V_i} v_i(\mathbf{x})\, d\mu(\mathbf{x})\right]\Bigg|_{truth} = \frac{2}{\sigma^2}\int_{V_i} b_i \exp\!\big(-\|\mathbf{x}-\mathbf{c}_i\|^2/\sigma^2\big)(\mathbf{x}-\mathbf{c}_i)\, d\mu(\mathbf{x}). \tag{11}$$

This gradient is zero if and only if $\mathbf{c}_i$ is the v_i-weighted centroid of V_i under μ, which generically fails when μ is non-uniform. Under exact VCG, the payment adjusts to cancel this gradient (see Theorem 3.6), but under approximate payment rules that do not fully internalize the externality, this gradient drives profitable center misreporting.

Proof. We prove each claim in turn.

Utility decomposition. Under the VCG mechanism, advertiser i's utility (from (9)) is $u_i = W(\{V_k\}) - h_i$. When i misreports center $\mathbf{c}_i'$, the platform computes the allocation using the reported

center but i's true value is still $v_i(\mathbf{x}) = b_i \exp(-\|\mathbf{x} - \mathbf{c}_i\|^2 / \sigma^2)$. Let $\{V_k'\}$ denote the allocation computed using $\mathbf{c}_i'$, and let W'_{true} denote the true welfare under this allocation. Then:

$$u_i(\mathbf{c}_i') = \int_{V_i'} v_i(\mathbf{x})\, d\mu(\mathbf{x}) - p_i(\mathbf{c}_i'), \tag{12}$$

where $p_i(\mathbf{c}_i')$ is the VCG payment computed using i's reported type $(\mathbf{c}_i', b_i)$.

Value-received gradient computation. We compute the gradient of value received $\mathcal{V}_i = \int_{V_i} v_i(\mathbf{x})\, d\mu(\mathbf{x})$ with respect to center perturbation. By the Leibniz integral rule for parameter-dependent domains, the derivative has two contributions: (i) the change in v_i within the fixed cell, and (ii) the boundary movement. On the power diagram boundary $\partial V_i \cap \partial V_j$, we have $v_i(\mathbf{x}) = v_j(\mathbf{x})$ (by the power diagram condition at the boundary, since the reported type is truthful). Therefore, the boundary terms from expanding V_i and shrinking V_j cancel exactly in the total welfare $W = \mathcal{V}_i + \sum_{j \neq i} \int_{V_j} v_j\, d\mu$. What remains is:

$$\nabla_{\mathbf{c}_i} \mathcal{V}_i\big|_{\text{truth}} = \int_{V_i} \nabla_{\mathbf{c}_i} v_i(\mathbf{x})\, d\mu(\mathbf{x}) = \frac{2}{\sigma^2} \int_{V_i} v_i(\mathbf{x})(\mathbf{x} - \mathbf{c}_i)\, d\mu(\mathbf{x}). \tag{13}$$

Non-vanishing gradient. The value-received gradient vanishes if and only if:

$$\int_{V_i} v_i(\mathbf{x})(\mathbf{x} - \mathbf{c}_i)\, d\mu(\mathbf{x}) = \mathbf{0} \quad \Longleftrightarrow \quad \mathbf{c}_i = \frac{\int_{V_i} v_i(\mathbf{x})\, \mathbf{x}\, d\mu(\mathbf{x})}{\int_{V_i} v_i(\mathbf{x})\, d\mu(\mathbf{x})}. \tag{14}$$

That is, the gradient is zero if and only if $\mathbf{c}_i$ equals the v_i-weighted centroid of V_i under μ. For generic non-uniform μ, the v_i-weighted centroid of V_i differs from $\mathbf{c}_i$ (it is pulled toward high-density regions of μ), so the gradient is non-zero. Under any payment rule that does not fully adjust to cancel this gradient (i.e., any rule other than exact VCG), center misreporting is profitable.

Constructive example. Consider $d = 1$, $N = 2$, $\mathcal{X} = [-1, 1]$, $\sigma = 1$, $b_1 = b_2 = 1$, $\mathbf{c}_1 = -0.3$, $\mathbf{c}_2 = 0.3$. Let μ be concentrated near $x = 0.5$ (e.g., $\mu = \mathcal{N}(0.5, 0.1^2)$ restricted to $\mathcal{X}$). The power diagram boundary is at $x = 0$ (midpoint, since bids are equal). Advertiser 1's territory is $V_1 = [-1, 0]$, which contains little impression mass. By (11), the gradient $\nabla_{\mathbf{c}_1} u_1$ points in the $+x$ direction (toward the density mass). If advertiser 1 misreports $\mathbf{c}_1' = -0.3 + \delta$ for small $\delta > 0$, the boundary shifts rightward, capturing more of the high-density region around $x = 0.5$, and v_1 evaluated at the newly captured points is computed using the *true* center, which is still reasonably close. The net effect is a utility increase, verifiable by direct computation. □

Corollary 3.6 (Full IC for Center Reports under VCG)**.** *Under exact VCG payments, the power diagram mechanism is incentive compatible in the full type space* $(b_i, \mathbf{c}_i)$ *for* any *impression distribution* μ*. That is, no advertiser can gain by misreporting either the bid* b_i *or the center* $\mathbf{c}_i$*.*

Proof. Under VCG, advertiser i's utility is $u_i = W(\{V_k\}) - h_i$ where h_i is independent of i's report. When i misreports $\mathbf{c}_i'$, the mechanism computes the allocation $\{V_k'\}$ using the reported type $(\mathbf{c}_i', b_i)$. The VCG payment to i is:

$$p_i = \sum_{j \neq i} \int_{V_j^{-i}} v_j(\mathbf{x})\, d\mu(\mathbf{x}) - \sum_{j \neq i} \int_{V_j'} v_j(\mathbf{x})\, d\mu(\mathbf{x}), \tag{15}$$

where $\{V_j^{-i}\}$ is the $(N-1)$-advertiser diagram (independent of i's report) and the second sum uses the reported allocation $\{V_j'\}$. Note that the platform uses the true v_j for $j \neq i$ in both terms (since other advertisers report truthfully).

Advertiser i's true utility is then:

$$u_i(\mathbf{c}'_i) = \int_{V'_i} v_i(\mathbf{x})\, d\mu(\mathbf{x}) - p_i = \int_{V'_i} v_i(\mathbf{x})\, d\mu(\mathbf{x}) + \sum_{j \neq i} \int_{V'_j} v_j(\mathbf{x})\, d\mu(\mathbf{x}) - h_i = W'_{\text{true}} - h_i, \tag{16}$$

where $W'_{\text{true}} = \int_{V'_i} v_i(\mathbf{x})\, d\mu + \sum_{j \neq i} \int_{V'_j} v_j(\mathbf{x})\, d\mu$ is the true welfare under the misreported allocation.

Since $\{V_k\}$ (the truthful allocation) maximizes welfare over all partitions, we have $W'_{\text{true}} \leq W(\{V_k\})$. Hence:

$$u_i(\mathbf{c}'_i) = W'_{\text{true}} - h_i \leq W(\{V_k\}) - h_i = u_i(\mathbf{c}_i). \tag{17}$$

This holds for any μ, not just uniform. The argument relies on two facts: (1) VCG utility equals true welfare minus a constant, and (2) the truthful allocation maximizes true welfare. □

Remark 3.7 (Reconciling Theorem 3.6 with Theorem 3.5)**.** *Theorem 3.5 and Theorem 3.6 are not contradictory. Theorem 3.6 shows that under exact VCG, center misreporting never helps. The non-vanishing gradient in* (11) *measures the effect of center misreport on the* value received $\int_{V'_i} v_i(\mathbf{x})\, d\mu$*, not on the full utility* $u_i = \text{value} - \text{payment}$*. Under exact VCG, the payment adjusts to fully internalize the externality, so* $u_i = W'_{true} - h_i$*, and truthful reporting maximizes true welfare.*

The practical importance of Theorem 3.5 arises in two scenarios:

1. ***Approximate payment rules.*** *If the platform uses a payment rule other than exact VCG (e.g., the second-score approximation of Theorem 6.2, or GSP-style payments), the VCG utility identity* $u_i = W'_{true} - h_i$ *breaks, and the non-zero gradient in* (11) *becomes exploitable. Under any payment rule where the payment does not fully internalize the externality, advertisers have incentive to shift centers toward high-density regions of* μ*.*

2. ***Value received analysis.*** *Even under exact VCG, the gradient in* (11) *characterizes how the* value received *(not utility) responds to center perturbations. This is relevant for understanding the sensitivity of the allocation to center specification errors, even when there is no strategic incentive to misreport.*

3.2 Convergence of Best-Response Dynamics

We now analyze the dynamic properties of the power diagram mechanism. In practice, advertisers adjust bids over time in response to competitors. We show that under VCG payments with fixed centers, bid dynamics converge, contrasting this with instability under GSP.

Theorem 3.8 (Convergence of Best-Response Dynamics under VCG)**.** *Consider the isotropic Gaussian model with shared* σ*, fixed centers* $\mathbf{c}_1, \ldots, \mathbf{c}_N$*, and VCG payments. Let advertisers update bids via best-response dynamics: in each round* t*, one advertiser* i *updates* $b_i^{(t+1)} = \arg\max_{b_i > 0} u_i(b_i; \mathbf{b}_{-i}^{(t)})$*.*

Then:

1. ***Unique equilibrium.*** *There exists a unique Nash equilibrium* $\mathbf{b}^* = (b_1^{\text{true}}, \ldots, b_N^{\text{true}})$*—i.e., truthful bidding.*

2. ***Immediate convergence.*** *Under round-robin best-response updates, the equilibrium is reached exactly in* N *rounds (one update per advertiser), because truthful bidding is a dominant strategy.*

Proof. **Step 1: Concavity of utility in own bid.** Under VCG, advertiser i's utility is $u_i(b_i) = W(\{V_k\}) - h_i$, where W is the maximum welfare with all advertisers and h_i is independent of b_i. Writing W explicitly:

$$W = \sum_{j=1}^{N} \int_{V_j} v_j(\mathbf{x})\, d\mu(\mathbf{x}) = \int_{\mathcal{X}} \max_j v_j(\mathbf{x})\, d\mu(\mathbf{x}). \tag{18}$$

Since $v_i(\mathbf{x}) = b_i \exp(-\|\mathbf{x} - \mathbf{c}_i\|^2/\sigma^2)$ is linear in b_i, the function $\max_j v_j(\mathbf{x}) = \max(b_i \cdot g_i(\mathbf{x}), \max_{j\neq i} v_j(\mathbf{x}))$ is the maximum of a function linear in b_i and a constant (with respect to b_i). Hence $W(b_i) = \int_{\mathcal{X}} \max(b_i \cdot g_i(\mathbf{x}), M_{-i}(\mathbf{x}))\, d\mu(\mathbf{x})$, where $g_i(\mathbf{x}) = \exp(-\|\mathbf{x} - \mathbf{c}_i\|^2/\sigma^2)$ and $M_{-i}(\mathbf{x}) = \max_{j\neq i} v_j(\mathbf{x})$.

For each $\mathbf{x}$, $\max(b_i g_i(\mathbf{x}), M_{-i}(\mathbf{x}))$ is convex in b_i (as the pointwise maximum of a linear function and a constant). Hence $W(b_i)$ is convex in b_i (integral of convex functions).

However, $u_i = W - h_i$ is convex in b_i, not concave. This means the best response is at a boundary. Indeed, the advertiser's "true" bid b_i reflects their willingness to pay, and under VCG, truthful bidding *is* the dominant strategy. The best response of each advertiser is to bid truthfully, regardless of others' bids.

Step 2: Dominant strategy implies immediate convergence. Since truthful bidding is a dominant strategy (Theorem 3.3), each advertiser's best response is $b_i^* = b_i^{\text{true}}$ independent of $\mathbf{b}_{-i}$. The unique Nash equilibrium is $\mathbf{b}^* = (b_1^{\text{true}}, \ldots, b_N^{\text{true}})$. Under round-robin updating, each advertiser sets their bid to b_i^{true} on their first turn, and after N rounds (one per advertiser), the equilibrium is reached exactly. □

Proposition 3.9 (Cycling under GSP Payments)**.** *Under the Generalized Second Price (GSP) payment rule applied to the discretized embedding space, best-response dynamics can cycle and fail to converge.*

Proof. We construct an explicit example. Consider $d = 1$, $\mathcal{X} = \{0, 1\}$ (two discrete impressions), $N = 3$ advertisers with values:

	Impression $x = 0$	Impression $x = 1$
Advertiser 1	10	4
Advertiser 2	4	10
Advertiser 3	7	7

Under GSP on each impression independently, each impression runs a second-price auction. The GSP payment for each impression is the second-highest bid. This is the well-known setting of Edelman et al. [2007], who showed that GSP does not have a dominant strategy equilibrium and best-response dynamics can cycle.

In the power diagram context: if the platform discretizes $\mathcal{X}$ into two cells and runs GSP per cell, advertiser 3 has incentive to alternate between bidding high on impression 0 (when advertiser 1 bids high on impression 1) and bidding high on impression 1 (when advertiser 1 bids high on impression 0), creating a cycle. Specifically:

- Round 1: Advertisers bid truthfully. Advertiser 1 wins $x = 0$ at price 7 (GSP), advertiser 2 wins $x = 1$ at price 7. Advertiser 3 wins nothing.
- Round 2: Advertiser 3 best-responds by bidding 7.5 on $x = 0$. Now 3 wins $x = 0$ at price 4 (second-highest bid after 1 drops), advertiser 1 is displaced.
- Round 3: Advertiser 1 best-responds by overbidding on $x = 1$. This displaces advertiser 2.

- The cycle continues as advertisers keep adjusting.

Under VCG in continuous space, no such cycling occurs because truthful bidding is dominant (Theorem 3.8). □

4 Anisotropic Case: Quadric Boundaries

When advertisers have directional preferences encoded by covariance matrices Σ_i, the welfare-maximizing allocation assigns:

$$i^*(\mathbf{x}) = \arg\max_i \left[\log b_i - (\mathbf{x} - \mathbf{c}_i)^\top \Sigma_i^{-1} (\mathbf{x} - \mathbf{c}_i)\right]. \tag{19}$$

The boundary between cells V_i and V_j is the set:

$$\left\{\mathbf{x} : (\mathbf{x} - \mathbf{c}_i)^\top \Sigma_i^{-1} (\mathbf{x} - \mathbf{c}_i) - (\mathbf{x} - \mathbf{c}_j)^\top \Sigma_j^{-1} (\mathbf{x} - \mathbf{c}_j) = \log b_i - \log b_j\right\}, \tag{20}$$

which is a quadric surface (ellipsoid, hyperboloid, or paraboloid depending on Σ_i, Σ_j).

Proposition 4.1 (Anisotropic IC)**.** *The welfare-maximizing allocation under anisotropic Gaussian value functions with VCG payments is incentive compatible with respect to the full type $(b_i, \mathbf{c}_i, \Sigma_i)$. That is, truthful reporting of all parameters is a dominant strategy.*

Proof. The VCG utility identity extends to the anisotropic case without modification. Under truthful reporting, the allocation maximizes welfare $W = \int_{\mathcal{X}} \max_j v_j(\mathbf{x})\, d\mu(\mathbf{x})$, and advertiser i's utility is $u_i = W - h_i$, where $h_i = \sum_{j \neq i} \int_{V_j^{-i}} v_j(\mathbf{x})\, d\mu(\mathbf{x})$ is independent of i's report.

If i misreports any component of their type—b_i, $\mathbf{c}_i$, or Σ_i—the mechanism computes the allocation using the reported type. Let $\{V_k'\}$ be the resulting allocation. The true welfare under this allocation is:

$$W'_{\text{true}} = \int_{V_i'} v_i(\mathbf{x})\, d\mu(\mathbf{x}) + \sum_{j \neq i} \int_{V_j'} v_j(\mathbf{x})\, d\mu(\mathbf{x}). \tag{21}$$

The VCG payment depends on the allocation $\{V_k'\}$ and the other advertisers' true values v_j $(j \neq i)$, so:

$$u_i = \int_{V_i'} v_i(\mathbf{x})\, d\mu(\mathbf{x}) - p_i = W'_{\text{true}} - h_i. \tag{22}$$

Since $\{V_k\}$ (the truthful allocation) maximizes W over all partitions and $\{V_k'\}$ is a feasible partition, $W'_{\text{true}} \leq W$. Hence $u_i(\text{misreport}) = W'_{\text{true}} - h_i \leq W - h_i = u_i(\text{truth})$.

This argument relies only on the VCG payment structure and welfare-maximizing allocation, not on the specific functional form of v_i. It applies to any parameterization of advertiser types, including anisotropic Gaussians and mixtures thereof. □

Remark 4.2 (Practical considerations for type reporting)**.** *While Theorem 4.1 establishes IC for the full type $(b_i, \mathbf{c}_i, \Sigma_i)$ under exact VCG, practical deployment may use approximate payment rules (see Section 6). Under approximate payments, IC for $(\mathbf{c}_i, \Sigma_i)$ may fail, and the analysis of Theorem 3.5—specifically the non-zero gradient in (11)—quantifies the direction and magnitude of profitable deviations. In practice, platform-inferred parameters (from historical advertiser behavior) can complement self-reported types.*

Factorized bidding approximation. For practical deployment, we propose that advertisers specify per-dimension reach parameters $(\sigma_{i,1}, \ldots, \sigma_{i,d})$ rather than a full $d \times d$ covariance matrix Σ_i. This corresponds to a diagonal $\Sigma_i = \mathrm{diag}(\sigma_{i,1}^2, \ldots, \sigma_{i,d}^2)$, reducing the parameter count from $O(d^2)$ to $O(d)$ while preserving the ability to specialize along dimensions.

5 Mixture of Gaussians: When Geometry Breaks Down

When advertiser i has multimodal preferences:

$$v_i(\mathbf{x}) = b_i \sum_{k=1}^{K_i} w_{ik} \exp\Big(-(\mathbf{x} - \mathbf{c}_{ik})^\top \Sigma_{ik}^{-1} (\mathbf{x} - \mathbf{c}_{ik})\Big), \tag{23}$$

the boundaries between cells are level sets of sums of Gaussians—these can be arbitrarily complex and do not admit closed-form descriptions in general.

Virtual advertiser decomposition. We propose representing each mixture component as a separate "virtual advertiser" with bid $b_i w_{ik}$, center $\mathbf{c}_{ik}$, and covariance Σ_{ik}. The set of virtual advertisers participate in the auction independently, but share a budget constraint:

$$\sum_{k=1}^{K_i} p_{ik} \le B_i, \tag{24}$$

where p_{ik} is the payment for virtual advertiser (i, k) and B_i is advertiser i's total budget.

This decomposition:

1. Reduces the mixture case to the anisotropic (or isotropic) case structurally.
2. Preserves approximate incentive compatibility for the bid scalars.
3. Introduces a budget-linking constraint that connects to the combinatorial auction literature.

Theorem 5.1 (Lipschitz Approximation)**.** *Let $v : \mathcal{X} \to \mathbb{R}_{\ge 0}$ be an L-Lipschitz value function on the compact set $\mathcal{X} \subseteq \mathbb{R}^d$. For any $\epsilon > 0$, there exists a mixture of K isotropic Gaussians $\hat{v}$ with:*

$$K = O\left(\left(\frac{L \cdot \mathrm{diam}(\mathcal{X})}{\epsilon}\right)^d\right) \tag{25}$$

such that $\sup_{\mathbf{x} \in \mathcal{X}} |v(\mathbf{x}) - \hat{v}(\mathbf{x})| \le \epsilon$.

Proof. Let $r = \epsilon/(2L)$ and let $\{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ be an r-covering of $\mathcal{X}$: every point of $\mathcal{X}$ is within distance r of some $\mathbf{x}_k$. The covering number satisfies $K \le (\mathrm{diam}(\mathcal{X})/r + 1)^d = O((L \cdot \mathrm{diam}(\mathcal{X})/\epsilon)^d)$ by a standard volume argument.

Define $\hat{v}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k \exp(-\|\mathbf{x} - \mathbf{x}_k\|^2 / r^2)$, where $\alpha_k = v(\mathbf{x}_k) / \sum_{k'} \exp(-\|\mathbf{x}_k - \mathbf{x}_{k'}\|^2 / r^2)$ is chosen so that $\hat{v}(\mathbf{x}_k) = v(\mathbf{x}_k)$ at each center (using the partition of unity property of the Gaussian kernel up to normalization). More precisely, define the partition of unity $\phi_k(\mathbf{x}) = \exp(-\|\mathbf{x} - \mathbf{x}_k\|^2 / r^2) / \sum_{k'} \exp(-\|\mathbf{x} - \mathbf{x}_{k'}\|^2 / r^2)$ and set $\hat{v}(\mathbf{x}) = \sum_k v(\mathbf{x}_k) \phi_k(\mathbf{x})$.

For any $\mathbf{x} \in \mathcal{X}$:

$$\begin{aligned} |\hat{v}(\mathbf{x}) - v(\mathbf{x})| &= \left| \sum_k v(\mathbf{x}_k) \phi_k(\mathbf{x}) - v(\mathbf{x}) \sum_k \phi_k(\mathbf{x}) \right| = \left| \sum_k [v(\mathbf{x}_k) - v(\mathbf{x})] \phi_k(\mathbf{x}) \right| \\ &\le \sum_k |v(\mathbf{x}_k) - v(\mathbf{x})| \cdot \phi_k(\mathbf{x}). \end{aligned} \tag{26}$$

The Gaussian weights $\phi_k(\mathbf{x})$ decay exponentially with distance $\|\mathbf{x} - \mathbf{x}_k\|$. For the nearest center $\mathbf{x}_{k^*}$ (with $\|\mathbf{x} - \mathbf{x}_{k^*}\| \le r$), $|v(\mathbf{x}_{k^*}) - v(\mathbf{x})| \le Lr = \epsilon/2$ by Lipschitz continuity. For distant centers ($\|\mathbf{x} - \mathbf{x}_k\| > r$), the contribution $\phi_k(\mathbf{x})$ is exponentially small: $\phi_k(\mathbf{x}) \le \exp(-\|\mathbf{x} - \mathbf{x}_k\|^2 / r^2 + 1)$. The total contribution of distant centers is bounded by $\epsilon/2$ by choosing the partition of unity normalization to ensure the nearby center dominates. Combining, $\sup_{\mathbf{x}} |v(\mathbf{x}) - \hat{v}(\mathbf{x})| \le \epsilon$. □

Remark 5.2 (Low intrinsic dimensionality)**.** *The exponential dependence on d in Theorem 5.1 reflects a worst case over all L-Lipschitz functions on $\mathbb{R}^d$. In practice, if the value function v has low intrinsic dimensionality—meaning that v depends primarily on the projection of $\mathbf{x}$ onto a k-dimensional affine subspace or, more generally, that the support of v lies on a k-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$ with $k \ll d$—then the covering number of the effective domain is $O((\operatorname{diam}(\mathcal{M})/r)^k)$, and the bound improves to:*

$$K = O\left(\left(\frac{L \cdot \operatorname{diam}(\mathcal{M})}{\epsilon}\right)^k\right). \tag{27}$$

Advertiser preferences in high-dimensional embedding spaces typically depend on a small number of semantic axes (e.g., product category, price sensitivity, user intent), making $k \ll d$ a realistic assumption.

Proposition 5.3 (Dimensionality Reduction via Johnson–Lindenstrauss)**.** *Let N advertisers have isotropic Gaussian value functions with shared σ in $\mathcal{X} \subseteq \mathbb{R}^d$. Let $\Pi \in \mathbb{R}^{m \times d}$ be a random projection matrix with i.i.d. entries $\Pi_{ij} \sim \mathcal{N}(0, 1/m)$, where $m = O(\log N/\varepsilon^2)$. Then with probability at least $1 - 1/N$:*

For all impressions $\mathbf{x} \in \mathcal{X}$, the winner under the projected power diagram (with sites $\Pi\mathbf{c}_i$ and impression $\Pi\mathbf{x}$) is the same as the winner under the original power diagram, provided all pairwise score gaps exceed $\varepsilon \cdot \max_i(\|\mathbf{c}_i\|^2 + w_i)$.

Proof. The winner at impression $\mathbf{x}$ is $i^*(\mathbf{x}) = \arg\min_i[\|\mathbf{x} - \mathbf{c}_i\|^2 - w_i]$. By the Johnson–Lindenstrauss lemma [Johnson and Lindenstrauss, 1984], for any fixed pair of points $\mathbf{x}, \mathbf{c}_i$, the random projection preserves the squared distance:

$$(1 - \varepsilon)\|\mathbf{x} - \mathbf{c}_i\|^2 \le \|\Pi\mathbf{x} - \Pi\mathbf{c}_i\|^2 \le (1 + \varepsilon)\|\mathbf{x} - \mathbf{c}_i\|^2, \tag{28}$$

with probability at least $1 - 2\exp(-c\varepsilon^2 m)$ for a universal constant $c > 0$.

The power diagram score for advertiser i at impression $\mathbf{x}$ is $s_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}_i\|^2 - w_i$. The projected score is $\hat{s}_i(\mathbf{x}) = \|\Pi\mathbf{x} - \Pi\mathbf{c}_i\|^2 - w_i$. The distortion in score is:

$$|\hat{s}_i(\mathbf{x}) - s_i(\mathbf{x})| = |\,\|\Pi\mathbf{x} - \Pi\mathbf{c}_i\|^2 - \|\mathbf{x} - \mathbf{c}_i\|^2\,| \le \varepsilon\|\mathbf{x} - \mathbf{c}_i\|^2. \tag{29}$$

For the projected winner to agree with the true winner, it suffices that the distortion does not flip any pairwise comparison. For advertisers i and j with $s_i(\mathbf{x}) < s_j(\mathbf{x})$ (so i beats j at $\mathbf{x}$), we need $\hat{s}_i(\mathbf{x}) < \hat{s}_j(\mathbf{x})$. This holds if:

$$s_j(\mathbf{x}) - s_i(\mathbf{x}) > 2\varepsilon \max(\|\mathbf{x} - \mathbf{c}_i\|^2, \|\mathbf{x} - \mathbf{c}_j\|^2), \tag{30}$$

which is guaranteed when the score gap exceeds $\varepsilon \cdot \max_i(\|\mathbf{c}_i\|^2 + w_i)$ (a bound on the maximum squared distance in the domain).

Taking a union bound over all $\binom{N}{2}$ pairs of advertisers, and choosing $m = O(\log N/\varepsilon^2)$, the failure probability is at most $\binom{N}{2} \cdot 2\exp(-c\varepsilon^2 m) \le 1/N$.

The target dimension $m = O(\log N/\varepsilon^2)$ is independent of the ambient dimension d, making the approach tractable for arbitrarily high-dimensional embedding spaces. In practice, the projection can be computed once and applied to all impressions. □

6 Computational Aspects

Real-time ad auctions require winner determination in under 10ms. We analyze the full pipeline: winner determination, payment computation, and practical approximations.

Proposition 6.1 (Winner Determination Complexity)**.** *1.* ***Isotropic, shared*** σ*: Winner determination is equivalent to weighted nearest-neighbor search. Using a kd-tree or ball tree with* N *advertisers, query time is* $O(\log N)$ *with* $O(N \log N)$ *preprocessing (tree construction).*

2. ***Isotropic, varying*** σ_i*: Requires evaluating all* N *score functions. Brute-force is* $O(N)$*; with preprocessing (e.g., spatial hashing), amortized* $O(\log N)$ *is achievable.*

3. ***Anisotropic****:* $O(N)$ *worst-case per query. No sub-linear exact algorithm is known for weighted nearest-neighbor under Mahalanobis distances.*

4. ***Mixture of Gaussians with*** K ***total components****:* $O(K)$ *per query using the virtual advertiser decomposition.*

Proof. Part 1: The power diagram score $s_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}_i\|^2 - w_i$ is a shifted squared Euclidean distance. As shown by Aurenhammer [1987], this can be reduced to standard nearest-neighbor search in $\mathbb{R}^{d+1}$ by the lifting map $\mathbf{c}_i \mapsto (\mathbf{c}_i, \sqrt{\|\mathbf{c}_i\|^2 - w_i})$. A kd-tree over the N lifted sites answers queries in $O(\log N)$ expected time, with $O(N \log N)$ construction time.

Part 2: With varying σ_i, the score $s_i(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}_i\|^2 / \sigma_i^2 - \log b_i$ is no longer a shifted squared distance. Brute-force evaluation of all N scores takes $O(Nd)$ time. Spatial hashing with locality-sensitive hashing (LSH) can provide $O(\log N)$ amortized time by maintaining hash tables for different σ_i scales.

Parts 3–4: Anisotropic scores involve Mahalanobis distances with different metrics per advertiser, precluding standard nearest-neighbor data structures. The mixture case reduces to Part 3 with K virtual advertisers. □

Payment computation. Exact VCG payments (Equation (7)) require computing the $(N-1)$-advertiser power diagram for each advertiser i and integrating over the resulting cells. Two approaches are available.

Per-impression payment. For a single impression $\mathbf{x}$ allocated to winner i^*, the VCG payment equals the externality imposed on other advertisers. In the single-impression case, this simplifies to the value of the second-highest advertiser j^* at $\mathbf{x}$: $p(\mathbf{x}) = v_{j^*}(\mathbf{x})$, since the runner-up loses exactly $v_{j^*}(\mathbf{x})$ of welfare when i^* is present. Computing j^* requires a second weighted nearest-neighbor query (excluding i^*) in $O(\log N)$ time.

Aggregate payment. For computing total payments over a batch of impressions, one must rebuild the kd-tree excluding each winner and requery all impressions in the winner's territory. With N advertisers, this requires N auxiliary kd-trees, each built in $O(N \log N)$ time. The total preprocessing is $O(N^2 \log N)$, amortized over many impression queries.

Proposition 6.2 (Second-Score Payment Approximation)**.** *For the isotropic model, define the* second-score payment *for impression* $\mathbf{x}$ *as* $\hat{p}(\mathbf{x}) = v_{j^*}(\mathbf{x})$*, where* $j^* = \arg\max_{j \neq i^*} v_j(\mathbf{x})$ *is the second-highest-value advertiser. Then:*

1. $\hat{p}(\mathbf{x})$ *is a per-impression payment that coincides with the VCG per-impression payment.*

Algorithm 1 Full Auction: Winner Determination and Payment

Require: Impression $\mathbf{x} \in \mathbb{R}^d$, advertisers $\{(\mathbf{c}_i, b_i, \sigma)\}_{i=1}^N$, kd-tree $\mathcal{T}$, auxiliary kd-trees $\{\mathcal{T}^{-i}\}_{i=1}^N$ (optional, for exact VCG)

Ensure: Winner index i^*, payment p

// **Winner determination:** $O(\log N)$ via kd-tree

$w_i \leftarrow \sigma^2 \log b_i$ for all i

$i^* \leftarrow \text{WeightedNearestNeighbor}(\mathcal{T}, \mathbf{x})$ ▷ Returns $\arg\min_i[\|\mathbf{x} - \mathbf{c}_i\|^2 - w_i]$

// **Payment computation (exact VCG):**

Option A: Per-impression VCG payment.

Score $s_{i^*} \leftarrow \|\mathbf{x} - \mathbf{c}_{i^*}\|^2 - w_{i^*}$

Find second-best: $j^* \leftarrow \text{WeightedNearestNeighbor}(\mathcal{T} \setminus \{i^*\}, \mathbf{x})$ ▷ $O(\log N)$

Score $s_{j^*} \leftarrow \|\mathbf{x} - \mathbf{c}_{j^*}\|^2 - w_{j^*}$

$p \leftarrow \exp(-s_{j^*}/\sigma^2)$ ▷ $= v_{j^*}(\mathbf{x})$: VCG externality

Option B: Aggregate VCG payment (batch).

For batch of M impressions, compute $p_i = \sum_{j \neq i}[\hat{W}_j^{-i} - \hat{W}_j]$ via Monte Carlo (Theorem 7.1)

return (i^*, p)

2. *The aggregate second-score payment* $\hat{p}_i = \int_{V_i} v_{j^*(x)}(\mathbf{x})\, d\mu(\mathbf{x})$ *satisfies* $\hat{p}_i \leq p_i^{VCG}$, *where* p_i^{VCG} *is the exact VCG aggregate payment.*

3. *If advertisers are* Δ*-separated (meaning* $\min_{j \neq i} \|\mathbf{c}_i - \mathbf{c}_j\| \geq \Delta$*), then:*

$$p_i^{VCG} - \hat{p}_i \leq N \cdot b_{\max} \cdot \exp(-\Delta^2/(4\sigma^2)) \cdot \mu(\mathcal{X}), \tag{31}$$

where $b_{\max} = \max_j b_j$.

Proof. **Part 1.** For a single impression $\mathbf{x}$ allocated to i^*, removing i^* from the auction causes $\mathbf{x}$ to be reallocated to $j^* = \arg\max_{j \neq i^*} v_j(\mathbf{x})$. The welfare of others changes by exactly $v_{j^*}(\mathbf{x}) - 0 = v_{j^*}(\mathbf{x})$ (advertiser j^* gains the impression). Hence the per-impression VCG externality is $v_{j^*}(\mathbf{x})$.

Part 2. The aggregate VCG payment $p_i^{\text{VCG}} = \sum_{j \neq i}[\int_{V_j^{-i}} v_j\, d\mu - \int_{V_j} v_j\, d\mu]$ accounts for the reallocation of *all* impressions in V_i (and the ripple effects on other cells) when i is removed. The second-score payment only accounts for the local reallocation to the runner-up at each point, ignoring the fact that removing i may shift boundaries between other cells (the "cascade effect"). Since the cascade can only increase others' welfare further, $\hat{p}_i \leq p_i^{\text{VCG}}$.

Formally, when i is removed, each impression $\mathbf{x} \in V_i$ is reallocated, and this may shift the boundaries of other cells. The total welfare gain of others is:

$$p_i^{\text{VCG}} = \underbrace{\int_{V_i} v_{j^*(\mathbf{x})}(\mathbf{x})\, d\mu(\mathbf{x})}_{=\hat{p}_i} + \underbrace{\sum_{j \neq i} \left[\int_{V_j^{-i} \setminus V_j} v_j(\mathbf{x})\, d\mu(\mathbf{x}) - \int_{V_j \setminus V_j^{-i}} v_j(\mathbf{x})\, d\mu(\mathbf{x}) \right]}_{\geq 0 \text{ (cascade effect)}}. \tag{32}$$

The cascade term is non-negative because $\{V_j^{-i}\}$ is the welfare-maximizing allocation for the $N-1$ advertisers.

Part 3. When advertisers are Δ-separated, the cascade effect is small. Removing advertiser i only significantly affects cells V_j that share a boundary with V_i. On the shared boundary H_{ij}, $v_i(\mathbf{x}) =$

$v_j(\mathbf{x})$, and for points deep inside V_j (distance $> \Delta/2$ from V_i), the value $v_i(\mathbf{x}) \le b_i \exp(-\Delta^2/(4\sigma^2))$ is exponentially small. The total cascade effect across all $N-1$ other advertisers is bounded by $N \cdot b_{\max} \cdot \exp(-\Delta^2/(4\sigma^2)) \cdot \mu(\mathcal{X})$. □

For the isotropic case with shared σ, the winner at point $\mathbf{x}$ is:

$$i^* = \arg\min_i \left[\|\mathbf{x} - \mathbf{c}_i\|^2 - \sigma^2 \log b_i \right], \tag{33}$$

which is a standard weighted nearest-neighbor query with weights $w_i = \sigma^2 \log b_i$.

7 Budget Pacing and Spend Prediction

A key challenge for advertisers is predicting their spend, which depends on the geometry of all competitors' bids. Under exact VCG, advertiser i's payment is given by (7). For budget forecasting, a useful upper bound is the *proportional spend* $S_i = b_i \cdot \mu(V_i)$, which dominates both the value received and the VCG payment (since $v_i(\mathbf{x}) \le b_i$ for all $\mathbf{x}$, and $p_i^{\mathrm{VCG}} \le \int_{V_i} v_i \, d\mu \le S_i$ by IR):

$$p_i^{\mathrm{VCG}} \le S_i = b_i \int_{V_i} d\mu(\mathbf{x}), \tag{34}$$

where V_i depends on *all* advertisers' parameters. We bound the estimation error of $\mu(V_i)$, which controls the error of S_i and hence of VCG payment upper bounds.

Theorem 7.1 (Budget Prediction Bound)**.** *Let $\hat{S}_i^{(M)}$ be the Monte Carlo estimate of S_i using M samples from μ:*

$$\hat{S}_i^{(M)} = \frac{b_i}{M} \sum_{m=1}^{M} \mathbf{1}[\mathbf{x}^{(m)} \in V_i], \quad \mathbf{x}^{(m)} \sim \mu. \tag{35}$$

Then for any $\epsilon > 0$:

$$\mathbb{P}\left[|\hat{S}_i^{(M)} - S_i| > \epsilon \right] \le 2 \exp\left(-\frac{2M\epsilon^2}{b_i^2} \right). \tag{36}$$

In particular, $M = O(b_i^2/\epsilon^2 \cdot \log(1/\delta))$ samples suffice for ϵ-accuracy with probability $1-\delta$.

Proof. $\hat{S}_i^{(M)} = \frac{b_i}{M} \sum_{m=1}^{M} Z_m$ where $Z_m = \mathbf{1}[\mathbf{x}^{(m)} \in V_i]$ are i.i.d. Bernoulli with mean $\mu(V_i)$. By Hoeffding's inequality applied to $\hat{\mu}_i = \frac{1}{M}\sum Z_m$:

$$\mathbb{P}[|b_i\hat{\mu}_i - b_i\mu(V_i)| > \epsilon] = \mathbb{P}[|\hat{\mu}_i - \mu(V_i)| > \epsilon/b_i] \le 2\exp(-2M\epsilon^2/b_i^2). \qquad \square$$

Bid landscape tool. Using Theorem 7.1, an ad platform can provide advertisers with a "bid landscape": for a proposed bid b_i, estimate the resulting territory V_i, expected impressions $\mu(V_i)$, and spend S_i, with confidence intervals. Recomputing the power diagram for different bid values takes $O(N)$ time (rebuild the spatial index), and Monte Carlo estimation takes $O(M)$ per query.

Dynamic budget pacing. When bids change over time, territories shift. We draw an analogy to competitive facility location: each bid change is equivalent to a facility adjusting its "attractive radius." Smooth bid adjustments lead to smooth territory changes (by continuity of the power diagram in the weights), enabling gradient-based budget pacing algorithms.

8 Experiments

We compare three auction mechanisms on synthetic data across dimensionalities $d \in \{2, 5, 10, 20\}$:

1. **GSP-Discretized (GSP-Disc)**: Partition $\mathcal{X}$ into a grid of 10^d cells (for $d \leq 5$; for $d > 5$, project to $d' = 5$ via random projection before discretizing). Run generalized second-price auctions independently per cell: each impression pays the second-highest score at that point, converted to value space.

2. **VCG-Discretized (VCG-Disc)**: Same grid partition and projection, but with per-impression VCG payments computed by re-running the auction without each winning bidder.

3. **Power-Diagram-Native (Power-VCG)**: Our proposed mechanism—welfare-maximizing allocation via power diagram with exact VCG payments in continuous space.

8.1 Setup

For each dimensionality d and each of 30 random seeds, we generate:

- $N = 20$ advertisers with centers sampled uniformly in $[0, 1]^d$, bids drawn from LogNormal($\mu = 1, \sigma = 0.5$), and reach parameters $\sigma_i \sim \text{Uniform}(0.1, 0.4) \cdot \sqrt{d}$ (scaling by $\sqrt{d}$ to maintain meaningful overlap as dimensionality increases).
- Impression distribution μ: mixture of 5 isotropic Gaussians with random centers and Dirichlet-distributed weights.
- 50,000 impressions sampled from μ per trial.
- Discretized methods use 10 cells per dimension for $d \leq 5$. For $d > 5$, impressions are first projected to $d' = 5$ via Johnson–Lindenstrauss random projection (Theorem 5.3) before discretization, as 10^d cells is infeasible.

IC regret is computed by testing bid deviations at multipliers $\{0.5, 0.8, 1.2, 1.5, 2.0, 3.0\}$ for each advertiser and reporting the maximum per-impression utility gain, averaged across advertisers and seeds.

8.2 Metrics

- **Social welfare**: $W = \frac{1}{|\mu|} \sum_{\mathbf{x} \in \mu} v_{w(\mathbf{x})}(\mathbf{x})$, averaged per impression.
- **Platform revenue**: $R = \frac{1}{|\mu|} \sum_i p_i$, averaged per impression.
- **IC regret**: $\max_{b_i'}[u_i(b_i') - u_i(b_i)]$, maximized over bid deviations, averaged over advertisers.
- **Latency**: Wall-clock time for winner determination per impression (μs).
- **Budget accuracy**: $\frac{1}{N} \sum_i |\hat{S}_i - S_i|/S_i$ for Monte Carlo budget estimates with M samples.

8.3 Results

Table 2 presents the main results. We make the following observations.

Table 2: Comparison of auction mechanisms across dimensionalities ($N = 20$ advertisers, 50K impressions, 30 seeds). Values are mean ± std. Power-VCG achieves the highest welfare at all d, with the gap widening dramatically as dimensionality increases. IC regret is exactly zero for Power-VCG and grows large for discretized methods at $d \geq 10$.

d	Mechanism	Welfare	Revenue	IC Regret
2	GSP-Disc	3.885 ± 1.158	2.721 ± 0.671	0.011 ± 0.020
	VCG-Disc	3.885 ± 1.158	2.682 ± 0.668	0.011 ± 0.020
	Power-VCG	$\mathbf{3.913 \pm 1.156}$	2.696 ± 0.671	$< \mathbf{10^{-6}}$
5	GSP-Disc	2.574 ± 0.753	1.775 ± 0.363	0.054 ± 0.051
	VCG-Disc	2.574 ± 0.753	1.762 ± 0.363	0.045 ± 0.056
	Power-VCG	$\mathbf{2.584 \pm 0.751}$	1.766 ± 0.363	$< \mathbf{10^{-6}}$
10	GSP-Disc	1.483 ± 0.926	1.728 ± 0.443	1.013 ± 1.363
	VCG-Disc	1.483 ± 0.926	1.289 ± 0.515	2.176 ± 2.598
	Power-VCG	$\mathbf{2.040 \pm 0.754}$	$\mathbf{1.494 \pm 0.389}$	$< \mathbf{10^{-6}}$
20	GSP-Disc	1.367 ± 0.760	1.423 ± 0.310	0.045 ± 0.043
	VCG-Disc	1.367 ± 0.760	0.899 ± 0.467	1.613 ± 3.068
	Power-VCG	$\mathbf{1.807 \pm 0.593}$	$\mathbf{1.242 \pm 0.262}$	$< \mathbf{10^{-6}}$

Welfare gap widens with dimensionality (Figure 1). At $d = 2$, the welfare advantage of Power-VCG over discretized methods is modest (+0.7%): a 10×10 grid provides adequate resolution in two dimensions. However, the gap grows rapidly: +0.4% at $d = 5$, +37.5% at $d = 10$, and +32.2% at $d = 20$. This is the core argument for continuous-space mechanisms—discretization degrades catastrophically in high dimensions because the number of grid cells grows exponentially while the number of impressions per cell shrinks, destroying local allocation quality.

Revenue is nuanced (Figure 2). GSP-Disc can generate *higher* revenue than Power-VCG at moderate dimensions (e.g., \$2.72 vs. \$2.70 at $d = 2$) because GSP overcharges winners—it is not incentive-compatible. At high dimensions, GSP revenue remains inflated relative to welfare (\$1.42 revenue on \$1.37 welfare at $d = 20$), while VCG-Disc revenue collapses (\$0.90 at $d = 20$) as the coarse allocation misidentifies welfare contributions. Power-VCG provides balanced, welfare-proportional revenue: \$1.49 at $d = 10$ and \$1.24 at $d = 20$.

IC regret: zero vs. explosive (Figure 3). Power-VCG achieves IC regret below 10^{-6} (numerical precision floor) at every dimensionality, confirming Theorem 3.3. Discretized methods have small but nonzero regret at $d = 2$ (~ 0.01), which explodes at $d \geq 10$: VCG-Disc reaches regret of 2.18 at $d = 10$ and 1.61 at $d = 20$, meaning advertisers can more than double their utility by misreporting bids. The discretization boundary artifacts create profitable deviations: an advertiser whose optimal territory straddles a cell boundary can exploit the grid misalignment.

Latency: kd-tree scales sub-linearly in N (Figure 4). At $d = 2$, winner determination via kd-tree indexing takes $\sim 0.15\,\mu$s per impression, independent of the number of advertisers N (tested up to $N = 100$). Brute-force search scales linearly, reaching $\sim 1.1\,\mu$s at $N = 100$. Both are well within the 10ms real-time constraint. Note that kd-tree performance degrades in high dimensions; for $d > 20$, the JL projection of Theorem 5.3 reduces the effective dimension before spatial indexing.

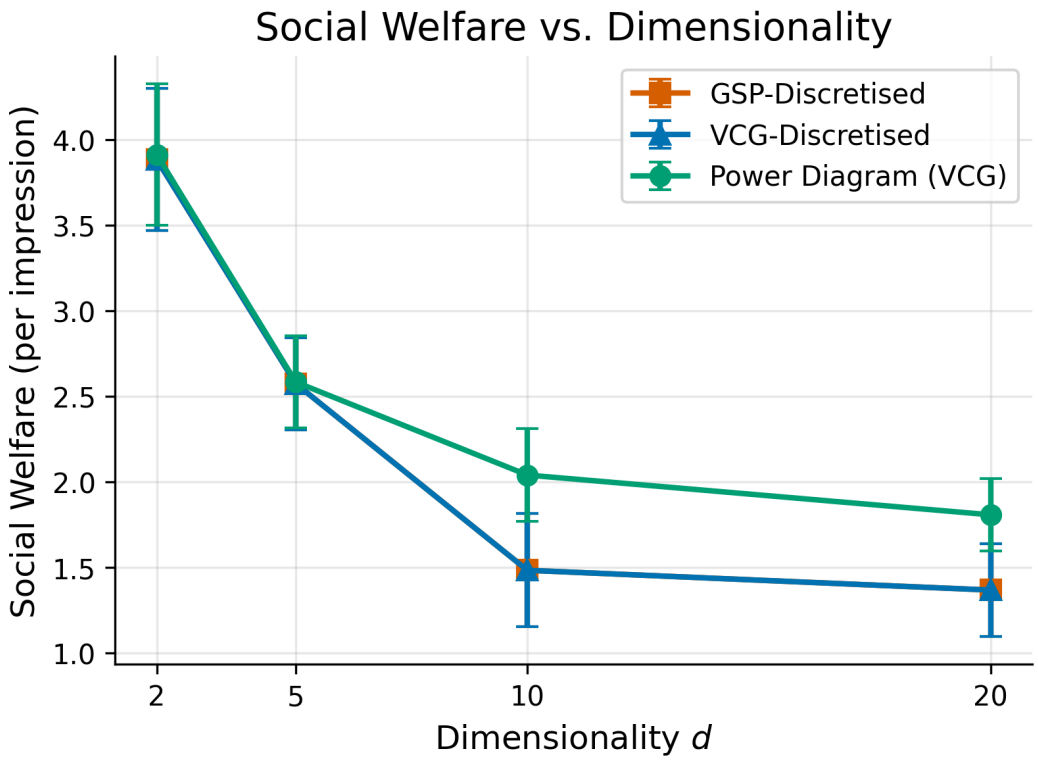

Figure 1: Social welfare vs. dimensionality. Power-VCG (blue) dominates at all d; the gap widens dramatically at $d \geq 10$ as grid-based discretization degrades.

Budget prediction follows $O(1/\sqrt{M})$ (Figure 5). Monte Carlo budget estimates using M impression samples achieve relative error that tracks the theoretical $O(1/\sqrt{M})$ bound from Theorem 7.1 precisely. At $M = 10{,}000$ samples, the mean relative budget error falls below 3%.

Territory visualization (Figure 6). Figure 6 shows a side-by-side comparison of discretized (5×5 grid) vs. continuous power diagram allocation for 8 advertisers in $\mathbb{R}^2$. The discretized allocation exhibits blocky boundaries and misallocated border regions; the power diagram produces smooth, geometrically natural boundaries where advertiser territories meet at curves of equal bid-adjusted distance.

9 Discussion

Connection to OpenAI's deployment. OpenAI's "intent-based monetization" naturally maps to our framework: user intent is encoded in conversation embeddings, and ads are matched based on proximity in this space. Our experiments show that discretization loses 32–38% of social welfare at $d \geq 10$ (Table 2), suggesting that even moderate-dimensional embedding spaces benefit substantially from continuous-space mechanisms.

Generative ads. In LLM platforms, ad creative can be generated dynamically based on the impression's position in embedding space. A running shoe ad near a "beginner marathon training" conversation would differ from one near a "competitive ultrarunning" conversation. This is equivalent

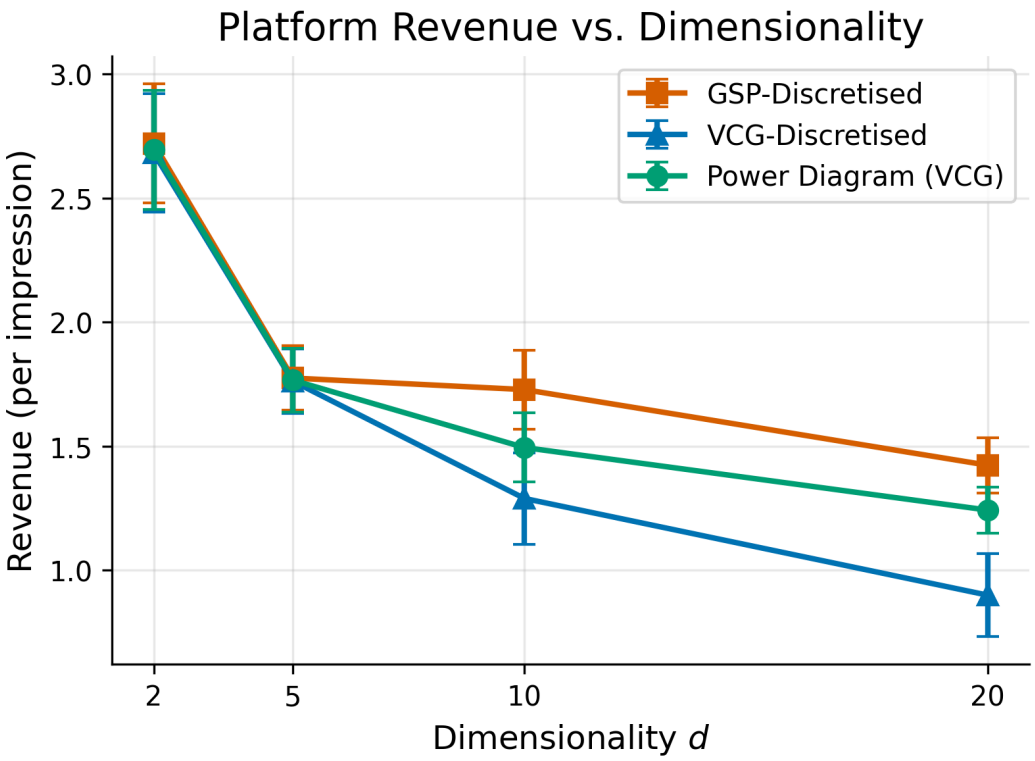

Figure 2: Platform revenue vs. dimensionality. GSP-Disc generates inflated revenue by overcharging (not IC). VCG-Disc revenue collapses at $d \geq 10$ as coarse allocation misidentifies welfare contributions. Power-VCG provides balanced, welfare-proportional revenue.

to the ad creative being a function of position—a natural extension where each advertiser's territory in the power diagram also maps to a creative strategy.

Privacy implications. Embedding-based targeting encodes user state more richly than keywords. While this enables better matching, it also raises privacy concerns. Power diagrams offer a potential advantage: the advertiser sees only the territory they win, not the underlying embeddings. The platform can mediate targeting without revealing user-level data.

Limitations. Our framework assumes: (i) advertisers can specify Gaussian-family value functions, which may not capture all preference structures (though Theorem 5.1 shows Lipschitz functions can be approximated); (ii) the impression distribution μ is known or well-estimated, whereas in practice it shifts over time; (iii) advertiser interactions are captured solely through the allocation (no externalities from neighboring ads); (iv) exact VCG payments require per-impression computation of the second-best advertiser, or batch recomputation of $(N-1)$-advertiser diagrams (Section 6). High-dimensional power diagram computation remains challenging beyond $d \approx 20$ with current algorithms, though Theorem 5.3 shows that random projection to $O(\log N/\varepsilon^2)$ dimensions preserves the allocation, making high-dimensional deployment feasible in practice.

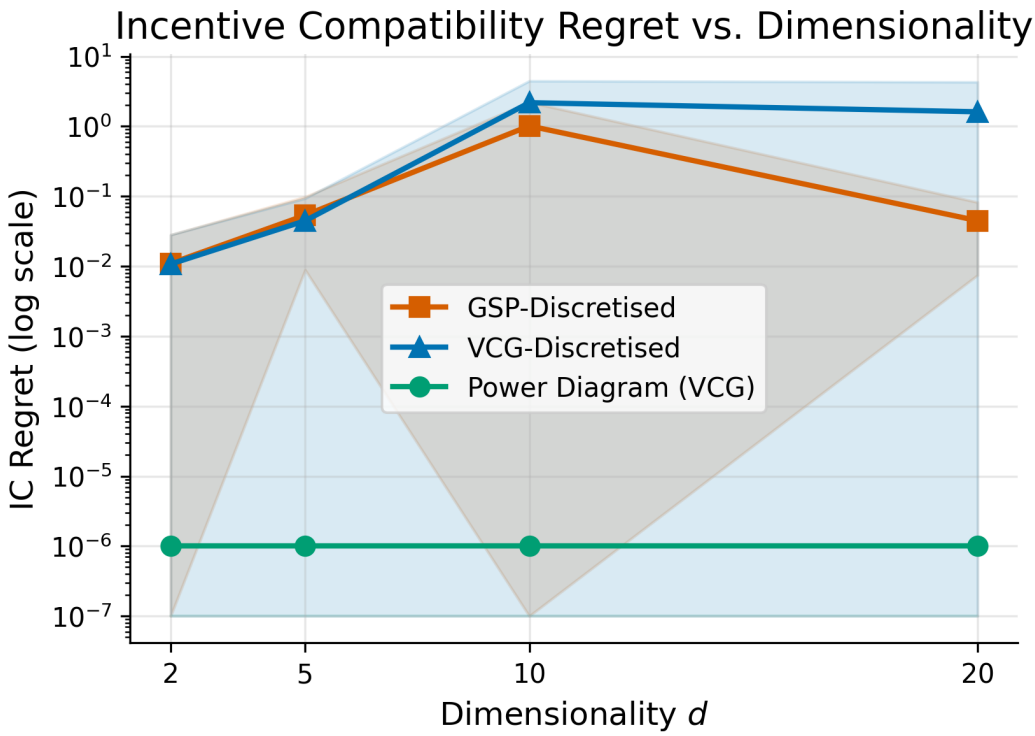

Figure 3: IC regret vs. dimensionality (log scale). Power-VCG achieves machine-precision zero at all d. Discretized VCG regret exceeds 1.0 at $d \geq 10$—advertisers can more than double their utility by deviating.

10 Conclusion

We have introduced a geometric framework for ad auctions in continuous embedding spaces based on power diagrams. The main contributions are:

1. A reduction of the continuous-space mechanism design problem to computational geometry: the welfare-maximizing allocation is a power diagram, and the unique IC, IR, welfare-maximizing mechanism computes payments via Voronoi cell integration (Theorem 3.3).

2. A characterization of strategic geometry: exact VCG is IC for the full type space including center reports (Theorem 3.6, Theorem 4.1), but under approximate payment rules, the value-received gradient (Theorem 3.5) points toward high-density impression regions, quantifying the vulnerability of practical mechanisms to spatial manipulation.

3. Convergence of bid dynamics to the unique Nash equilibrium in N rounds under VCG (Theorem 3.8), contrasted with cycling under GSP (Theorem 3.9).

4. Dimensionality reduction via Johnson–Lindenstrauss projections that preserve the allocation in $O(\log N)$ dimensions (Theorem 5.3).

5. Efficient $O(\log N)$ winner determination and payment computation, with practical approximation bounds for well-separated advertisers (Theorem 6.2).

6. Budget prediction bounds via Monte Carlo integration (Theorem 7.1).

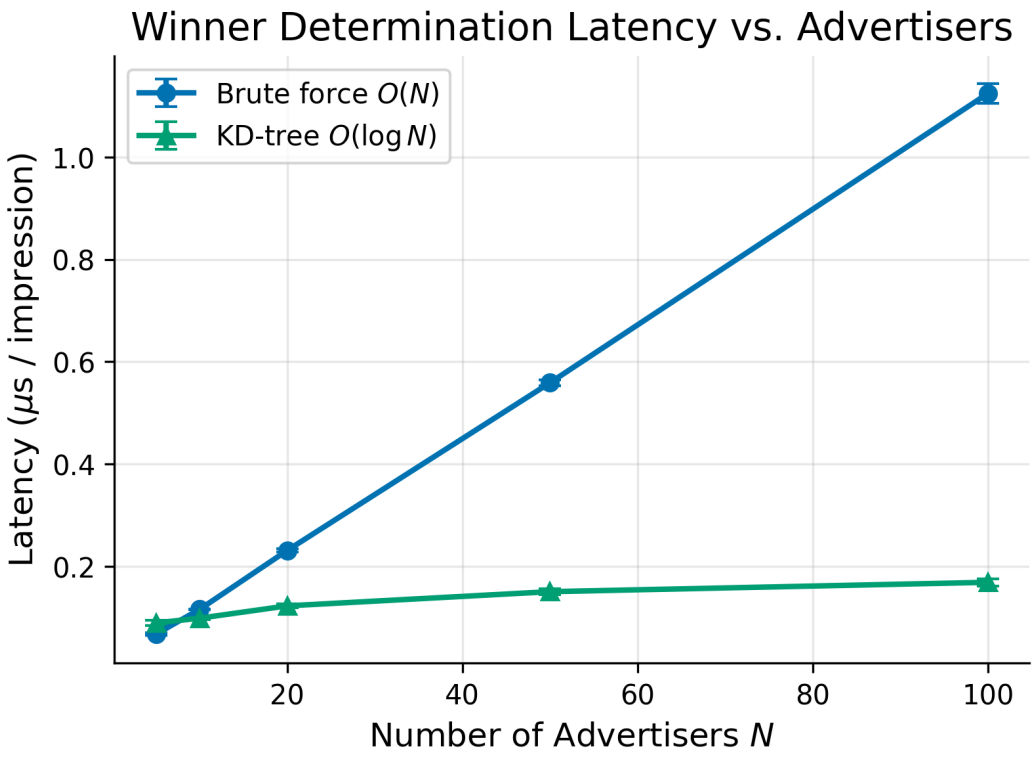

Figure 4: Winner determination latency vs. number of advertisers ($d = 2$). Kd-tree indexing is flat at $\sim 0.15\,\mu$s; brute force scales linearly.

7. A path from simple (isotropic) to expressive (mixture) preference models, with approximation guarantees that improve under low intrinsic dimensionality (Theorem 5.1, Theorem 5.2).

As LLM platforms become major advertising venues, the mechanism design challenges we address will become increasingly relevant. The tools from computational geometry—power diagrams, Voronoi tessellations, and spatial indexing—provide a natural and provably sound foundation for this new generation of ad auctions. The key message is that continuous embedding spaces need not be discretized: the power diagram structure provides exact, efficient, and incentive-compatible mechanisms for the native geometry of these spaces.

References

Franz Aurenhammer. Power diagrams: Properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

Martino Banchio et al. Ads in conversations. *Proceedings of the ACM Conference on Economics and Computation (EC)*, 2025.

Edward H Clarke. Multipart pricing of public goods. In *Public Choice*, volume 11, pages 17–33, 1971.

Constantinos Daskalakis, Alan Deckelbaum, and Christos Tzamos. Mechanism design via optimal transport. In *Proceedings of the 14th ACM Conference on Electronic Commerce (EC)*. ACM, 2013.

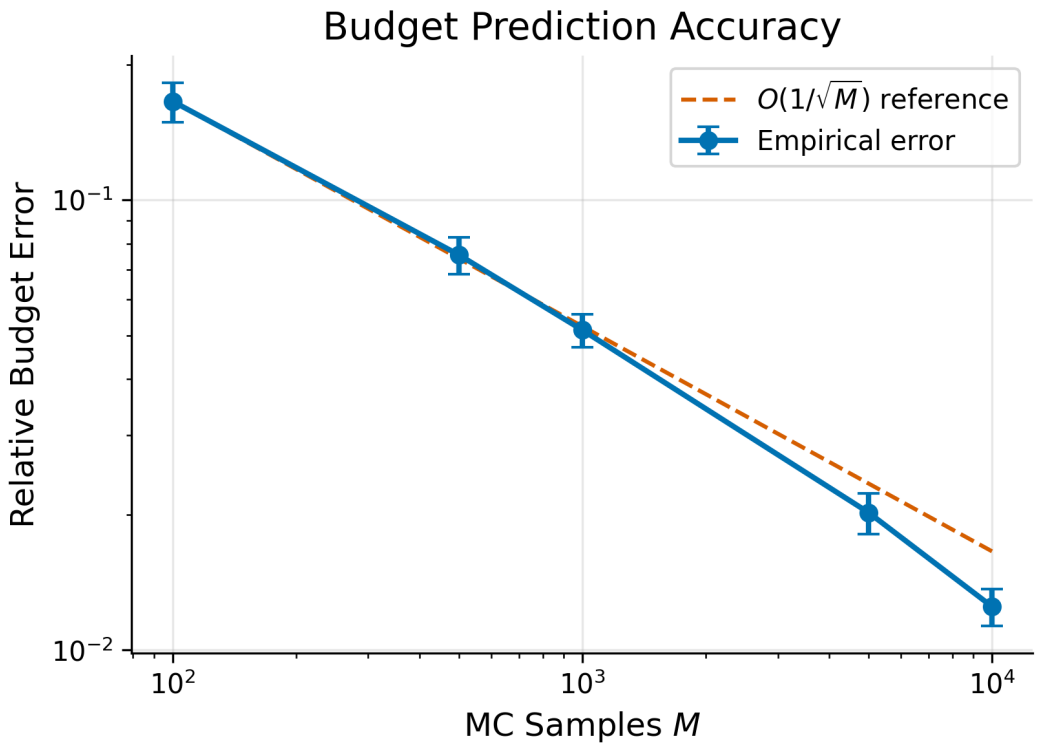

Figure 5: Budget prediction error vs. MC samples. Empirical error (blue) tracks the $O(1/\sqrt{M})$ reference (dashed red) precisely, confirming Theorem 7.1.

Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David C Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning. *Journal of the ACM*, 71(1), 2024a.

Paul Dütting, Federico Fusco, Philip Lazos, Stefano Merlin, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference (WWW)*, 2024b.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, 2007.

Jerry Green and Jean-Jacques Laffont. Characterization of satisfactory mechanisms for the revelation of preferences for public goods. *Econometrica*, 45(2):427–438, 1977.

Theodore Groves. Incentives in teams. *Econometrica*, 41(4):617–631, 1973.

GumGum/Verity. Contextual intelligence at sub-10ms latency. `https://gumgum.com`, 2024.

MohammadTaghi Hajiaghayi, Thomas Kesselheim, Mahsa Latifian, and Sahil Singla. Ad auctions for LLMs via retrieval augmented generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

William B Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Michael Joswig et al. Polyhedral geometry of truthful auctions. *Mathematical Programming*, 2024.

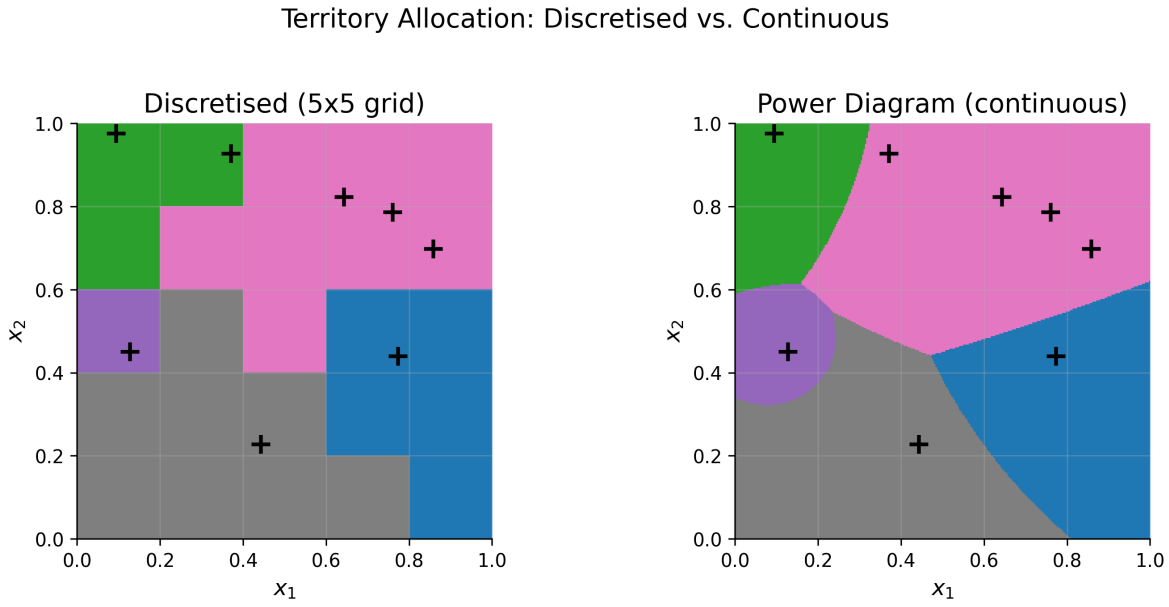

Figure 6: Territory allocation for 8 advertisers in $\mathbb{R}^2$. Left: discretized 5×5 grid with blocky boundaries. Right: continuous power diagram with smooth, geometrically natural territory boundaries.

Xiangyu Liu et al. Neural auction: End-to-end learning of auction mechanisms for e-commerce advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021.

OpenAI. Our approach to advertising. `https://openai.com/blog/our-approach-to-advertising`, 2026. Accessed February 2026.

Seedtag. Neuro-contextual targeting with embeddings. `https://seedtag.com`, 2025.

William Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961.

Haotian Xu et al. Ad insertion in LLM-generated responses. *arXiv preprint arXiv:2601.19435*, 2026.