FM PARAMETER ESTIMATION WITH LOW-ORDER RATIONAL CONSTRAINTS ON WASSERSTEIN LOSS LANDSCAPE

Ryoya Tabata
Research and Development Division
Yamaha Corporation
Hamamatsu, Japan
ryoya.tabata@music.yamaha.com

Masaki Iwaya
Digital Musical Instruments Division
Yamaha Corporation
Hamamatsu, Japan
masaki.iwaya@music.yamaha.com

Kazunobu Kondo
Research and Development Division
Yamaha Corporation
Hamamatsu, Japan
kazunobu.kondo@music.yamaha.com

ABSTRACT

Frequency modulation (FM) synthesis has been widely used in music production and sound design due to its ability to generate rich timbres with few control parameters. However, estimating the frequency parameters from a target sound remains challenging because different parameter configurations can yield similar spectra, creating numerous local minima in the loss landscape. In this paper, we analyze the Wasserstein distance loss landscape for two-operator FM synthesis under practical FFT-based spectral representations and show that it exhibits non-differentiable ridges at rational frequency ratios, arising from negative-frequency folding and spectral ordering transitions. Exploiting this structure, we propose a constrained gradient-based optimization strategy that constrains the frequency ratio in each optimization run to an interval bounded by consecutive low-order rational ratios and retains the lowest-loss candidate across intervals. Experimental results from controlled ablations show that maintaining the constraint throughout optimization improves reliability over random initialization and initialization-only constraints, particularly for more complex spectra at higher modulation indices.

1. INTRODUCTION

Frequency modulation (FM) synthesis [1, 2] is a widely used sound synthesis technique capable of generating rich and complex timbres with few control parameters. However, the inverse problem, i.e., estimating FM parameters from a target sound, remains challenging due to the non-convex nature of the loss landscape arising from the nonlinear relationship between FM parameters and the resulting spectrum.

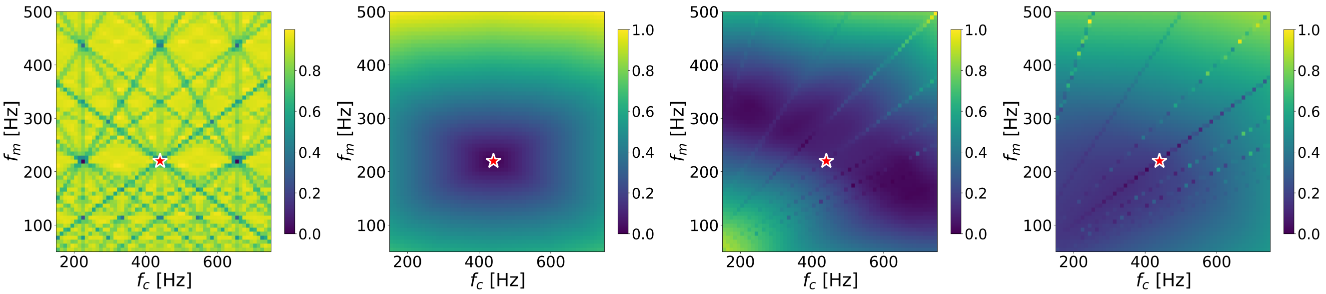
Early approaches employed genetic algorithms (GA) [3] and other evolutionary strategies [4] to search the parameter space. While effective at escaping local minima, these methods require substantial computational cost and provide no guarantee of convergence to the global optimum, highlighting the need for more efficient optimization methods.

Model-based signal-processing approaches have also addressed related sinusoidal and FM parameter estimation problems. For example, time-frequency ridge-based methods estimate instantaneous frequency and amplitude trajectories of multicomponent AM–FM signals from STFT representations [5], while distribution-derivative methods have been adapted to parametric FM sinusoids by exploiting the sideband structure of the FM spectrum [6]. These methods provide direct estimators for model-specific quantities, such as modal ridges or sideband-related FM parameters, in contrast to the data-driven and optimization-based approaches discussed next.

Deep learning approaches have recently gained attention. Encoder based methods such as Sound2Synth [7] learn a direct mapping from audio features to synthesizer parameters, achieving fast inference but requiring large training datasets and often struggling with out-of-distribution sounds. Flow based generative models [8, 9] can capture the multi-modal nature of the inverse problem by estimating parameter distributions rather than point estimates, though they may trade reconstruction accuracy for diversity.

Differentiable synthesis approaches like DDX7 [10] enable gradient-based optimization through automatic differentiation. However, these methods face inherent difficulties with frequency parameters due to the oscillatory nature of spectral losses. As noted by Turian and Henry [11], standard spectral losses lack stable and informative frequency gradients. Hayes et al. [12] addressed this for sinusoidal models by proposing a complex exponential surrogate with Wirtinger derivatives, demonstrating that gradient-based frequency estimation becomes tractable when the loss landscape is appropriately transformed. While their surrogate enables end-to-end learning for signals composed of multiple sinusoidal partials, FM synthesis presents additional challenges: the carrier–modulator interaction creates intricate spectral patterns governed by Bessel functions, leading to even more complex loss landscapes with dense local minima. In our preliminary experiments, we extended their surrogate approach to 2-operator FM synthesis but found that optimization converges only when initialized sufficiently close to the true parameters, otherwise becoming trapped in local minima.

Optimal transport distances offer an effective alternative for spectral matching. Unlike pointwise spectral losses such as spectral mean squared error (spectral MSE), optimal transport measures the cost of transporting spectral mass between distributions, providing gradients that reflect spectral proximity even when peaks do not overlap. Torres et al. [13] proposed spectral optimal transport (SOT) losses for differentiable DSP, demonstrating improved pitch estimation in harmonic synthesizers by minimizing the displacement of spectral energy. Matsubayashi [14] demonstrated that Wasserstein distance based losses yield smoother optimization landscapes for FM synthesis compared to spectral MSE, where the latter produces dense oscillatory patterns that trap gradient-based optimizers. While these works highlight the potential of optimal transport for inverse synthesis, they do not provide a systematic analysis of the loss landscape structure specific to FM synthesis.



(a) Spectral MSE: Bessel **(b)** W_1: Bessel **(c)** W_1: Bessel (folded) **(d)** W_1: FFT

Figure 1: Loss landscapes for carrier frequency $f_c = 440$ Hz, modulator frequency $f_m = 220$ Hz, $\beta = 3$; the red star indicates the target parameters. (a) Spectral MSE: Bessel spectrum. (b) W_1: Bessel spectrum. (c) W_1: Bessel spectrum with negative-frequency folding. (d) W_1: FFT spectrum.

Existing studies on differentiable FM synthesis typically avoid joint optimization of the frequency parameters: either the carrier frequency f_c is determined from an externally estimated pitch while the frequency ratio $\omega = f_m/f_c$, where f_m denotes the modulator frequency, is fixed to a preset value, or f_c is assumed known while ω is optimized. In both cases, f_c and ω are not optimized simultaneously, limiting applicability to scenarios where at least one frequency parameter is determined a priori.

In this paper, we propose a gradient-based optimization strategy that enables simultaneous optimization of carrier and modulator frequencies without relying on external pitch trackers. Frequency estimation is a challenging aspect of FM parameter recovery, as the loss landscape is highly sensitive to frequency parameters and exhibits multiple local minima. We focus on 2-operator FM synthesis as a concrete, controlled setting for analysis and experimental evaluation, treating the modulation index β as known within each estimation run while optimizing f_c and f_m. This choice is motivated by the structure analyzed in Section 4: the folding and ordering-transition boundaries occur at rational values of the frequency ratio and are independent of β, whereas β changes the relative Bessel sideband weights and hence the prominence of the ridges. Holding β fixed within each run lets us evaluate the proposed frequency-ratio constraint in isolation from changes in the sideband-weight distribution; by evaluating several β values, we still test spectra of varying complexity. Throughout, we operate on normalized power spectra (squared FFT magnitudes); the overall amplitude A therefore does not affect the loss, and both A and phase are outside the scope of this magnitude-based formulation. To develop this approach, we analyze the loss landscape of the Wasserstein distance for FM synthesis under practical FFT-based spectral representations. We show that the loss exhibits non-differentiable ridges at rational frequency ratios $\omega = m/n$, arising from two mechanisms: frequency folding creates V-shaped cusps at $\omega = 1/n$ where the n-th lower sideband crosses zero frequency, while ordering transitions at other rational ratios cause discontinuities when spectral components coincide or exchange their ordering. The folding of negative-frequency sidebands significantly alters the loss landscape structure, introducing nondifferentiable ridges and local minima that can trap or mislead gradient-based optimizers. Based on this analysis, we constrain the frequency ratio within intervals bounded by consecutive low-order rational boundaries, solving multiple constrained optimization problems and selecting the best solution. Controlled ablations with random and initialization-only variants isolate the effect of maintaining the constraint; experiments show improved reliability and reduced worst-case error.

2. FM SYNTHESIS

In two-operator FM synthesis (2OP-FM) [1, 2], a carrier oscillator is frequency-modulated by a modulator oscillator. The output signal is given by:

$$y(t) = A \sin\left(2\pi f_c t + \beta \sin(2\pi f_m t)\right), \tag{1}$$

where A is the amplitude, f_c is the carrier frequency, f_m is the modulator frequency, and β is the modulation index.

Using the Jacobi-Anger expansion, the signal can be decomposed into a sum of sinusoidal components:

$$y(t) = A \sum_{n=-\infty}^{\infty} J_n(\beta) \sin\left(2\pi(f_c + n f_m)t\right), \tag{2}$$

where $J_n(\beta)$ is the Bessel function of the first kind of order n. This reveals that the FM spectrum consists of spectral components at frequencies $f_c + n f_m$ with amplitudes proportional to $|J_n(\beta)|$.

The frequency ratio $\omega = f_m/f_c$ determines the harmonic structure. When ω is rational, i.e., $\omega = m/n$ for positive integers m and n, the spectrum is periodic and can be expressed as integer multiples of a fundamental frequency. When ω is irrational, the spectrum is non-periodic and the partials are inharmonic.

Chowning [1] describes how negative-frequency components are reflected at 0 Hz and added to positive-frequency components, increasing spectral complexity and richness and enabling wide, flexible control over timbre. This folding effect is central to the loss-landscape analysis in Section 4.

3. OPTIMAL TRANSPORT

The Wasserstein-p distance [15] between two probability measures μ, ν on a metric space (X, d) is defined as:

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times X} d(x,y)^p \, d\pi(x,y) \right)^{1/p}, \tag{3}$$

where $\Pi(\mu, \nu)$ denotes the set of all transport plans (couplings) with marginals μ and ν, and $p \geq 1$. For $p = 1$, this reduces to the Earth Mover's Distance, which represents the minimum cost to transport mass from μ to ν.

In the one-dimensional case ($X = \mathbb{R}$, $d(x, y) = |x - y|$), optimal transport preserves order: crossing transports always incur extra cost. For $x_1 < x_2$ and $y_1 < y_2$:

$$|x_1 - y_2| + |x_2 - y_1| \geq |x_1 - y_1| + |x_2 - y_2|. \tag{4}$$

In one dimension, this monotonicity property characterizes the co-monotone transport plan, which is optimal for convex displacement costs and, for discrete measures, is obtained by cumulative mass matching over the ordered supports [16].

The cumulative distribution function (CDF) is defined for a measure μ as $F_\mu(t) := \mu((-\infty, t])$. The difference $F_\mu(t) - F_\nu(t)$ represents the excess (or deficit) of mass to the left of t, which must cross point t during transport. Integrating the magnitude of this imbalance over all t yields the well-known one-dimensional closed-form expression that coincides with the $p = 1$ specialization of (3) for $X = \mathbb{R}$ and $d(x, y) = |x - y|$ [16, Prop. 2.17; see also Sec. 5.1]:

$$W_1(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)|\, dt. \tag{5}$$

For discrete distributions $\mu = \sum_i a_i \delta_{x_i}$ and $\nu = \sum_j b_j \delta_{y_j}$, the CDFs become step functions. Let $S = \{x_i\}_i \cup \{y_j\}_j$ be the union of the two support-point sets, and write its elements in increasing order as $t_0 < t_1 < \cdots < t_K$. Then W_1 reduces to:

$$W_1(\mu, \nu) = \sum_{k=0}^{K-1} |F_\mu(t_k) - F_\nu(t_k)| \cdot (t_{k+1} - t_k). \tag{6}$$

This formulation makes W_1 analytically tractable for discrete spectra. When the support points $\{x_i\}$ depend continuously on a parameter ω, $W_1(\omega)$ is continuous. Moreover, within any region where the ordering of support points remains fixed, $W_1(\omega)$ inherits the smoothness of the support point functions (piecewise linear when the dependence is linear). Non-differentiability occurs at parameter values where either the ordering changes or the support points themselves are non-differentiable.

In this work, we adopt $p = 1$ (W_1) as our loss function. The W_1 distance admits the closed-form CDF expression (Eq. (5)), which yields piecewise-linear dependence on the spectral positions and facilitates the analytical characterization of non-differentiable ridges presented in Section 4. We verified empirically that W_2 produces qualitatively identical loss landscape structure, including the same ridge patterns at rational frequency ratios, confirming that the ridge phenomenon is not specific to W_1. However, W_1 gave more stable optimization than W_2 in this setting. In one dimension, this metric choice amounts to the familiar L_1-versus-L_2 norm distinction, applied to quantile (transport) displacements rather than to pointwise spectral amplitudes.

4. LOSS LANDSCAPE ANALYSIS

To visualize the behavior of the Wasserstein distance as a function of FM synthesis parameters, we compute the W_1 distance between the spectrum of a reference FM signal (with $f_c = 440$ Hz, $f_m = 220$ Hz, $\beta = 3$, and sampling rate $f_s = 44\,100$ Hz) and spectra generated by varying the carrier frequency f_c and modulator frequency f_m. Unless otherwise noted, all loss-landscape illustrations in Figs. 1–5 use this same analysis setting.

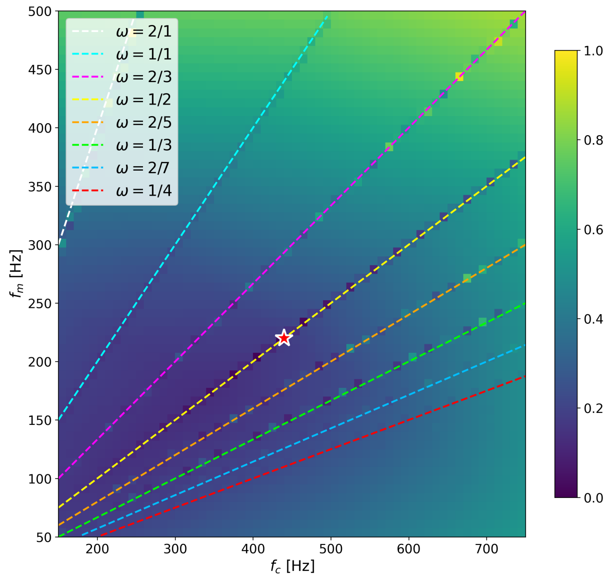


Figure 2: W_1 distance landscape (same as Fig. 1d) with rational frequency-ratio lines $\omega = m/n$ overlaid. The ridges align with these rational boundaries.

Before comparing these landscapes, recall from Section 2 that real-valued FM spectra fold negative-frequency components onto the positive-frequency axis. Thus, a lower sideband at $f_c + n f_m = f_c(1 + n\omega) < 0$ is observed at $\tilde{f}_n(\omega) = |f_c(1 + n\omega)|$ in the FFT power spectrum. This folding changes the support locations and ordering seen by the one-dimensional W_1 distance.

Fig. 1 compares loss landscapes under different spectral representations. Panel (a) shows spectral MSE between Bessel-expanded spectra without negative-frequency folding; the absence of transport structure leads to numerous local minima and near-zero gradients away from exact peak alignment. Panels (b)–(d) show W_1 distances under three representations: (b) Bessel spectrum without folding (full frequency axis $[-f_s/2, f_s/2]$), (c) Bessel spectrum with negative-frequency components folded (reflected) onto the positive axis $[0, f_s/2]$, and (d) actual FFT power spectrum. While (b) is smooth, it does not reflect the real-valued FFT behavior where negative frequencies fold onto positive ones. Panel (c) incorporates folding analytically, revealing diagonal ridges that also appear in the FFT-based landscape (d). The close agreement between (c) and (d) confirms that the Bessel expansion with folding captures the essential structure of the practical FFT-based loss.

Fig. 2 overlays lines of constant frequency ratio $\omega = m/n$ on the FFT-based loss landscape, confirming that the observed ridges align with rational frequency ratios.

The sideband-level mechanism behind these ridges is illustrated in Figs. 3 and 4. Fig. 3 shows a folded FM spectrum at $\omega = 0.45$ and $\beta = 3$, where the folded lower sidebands (dashed) appear at positions distinct from, but close to, the upper sidebands. As ω varies, each folded sideband traces a V-shaped curve with a cusp at the rational value $\omega = 1/|n|$ (Fig. 4). This folding is the key mechanism that transforms the smooth W_1 landscape of Fig. 1(b) into the ridged landscape of Figs. 1(c)–(d).

The loss landscape reveals several key properties:

- **Target minimum:** The W_1 distance reaches zero at the true parameters (f_c^*, f_m^*), confirming that the target parameters are represented exactly in the plotted landscape.

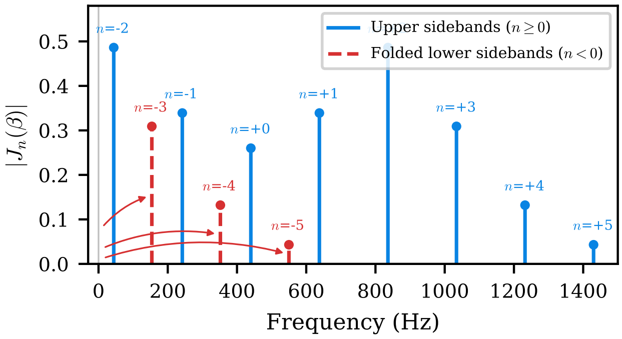


Figure 3: FM spectrum after negative-frequency folding (f_c = 440 Hz, $\omega = 0.45$, $\beta = 3$). Solid: upper sidebands ($n \geq 0$); dashed: folded lower sidebands ($n < 0$).

- **Non-differentiable ridges:** Diagonal ridges emanate from the origin along lines of rational frequency ratio $\omega = f_m/f_c = m/n$. These arise from two mechanisms:
  1. *Frequency folding:* The folded frequency $\tilde{f}_{-n}(\omega) = |f_c(1-n\omega)|$ is a V-shaped function with a non-differentiable cusp at $\omega = 1/n$, where the derivative jumps by $2nf_c$. This discontinuity propagates directly to W_1 via the chain rule.
  2. *Ordering transitions:* Two folded sidebands $\tilde{f}_{k_1}$ and $\tilde{f}_{k_2}$ coincide when $\omega = 2/|k_1+k_2|$ (e.g., $(k_1, k_2) = (1,-4), (1,-6), (-3,-4)$ give $\omega = 2/3, 2/5, 2/7$), causing a reordering of spectral components. Since the W_1 distance depends on spectral ordering through the CDF (Eq. (6)), such reorderings produce gradient discontinuities.

  Both effects produce non-differentiable boundaries. Low-order rationals (e.g., $2/3, 2/5, 1/2$) exhibit stronger ridges because they involve low-order sidebands with larger Bessel coefficients, while high-order rationals produce negligible ridges due to the rapid decay of $|J_n(\beta)|$ with $|n|$. Consequently, the loss landscape is effectively smooth between consecutive low-order rational boundaries, so constraining the frequency ratio to a single such region removes the competing minima that lie in other wedges.

Beyond the ridges, the loss landscape contains competing local minima, located both along the rational-ratio ridges and within the wedge interiors away from any ridge. In practice, the non-differentiability mainly marks the region boundaries, while the optimization difficulty comes from the competing basins and local minima organized by them.

As can be inferred from the underlying mechanism, the ridge locations depend primarily on the predicted parameters rather than the target; consequently, the same ridge structure emerges regardless of the target spectrum.

An important observation is that while the positions of these ridges are determined purely by the frequency-ratio geometry and are independent of the modulation index β, the prominence of each ridge depends strongly on β through the Bessel coefficients of the involved sidebands. Fig. 5 compares the W_1 loss landscape across several values of β. At $\beta = 1$, only sidebands up to order $|n| \approx 2$ carry significant energy (Carson's rule), so only the lowest-order

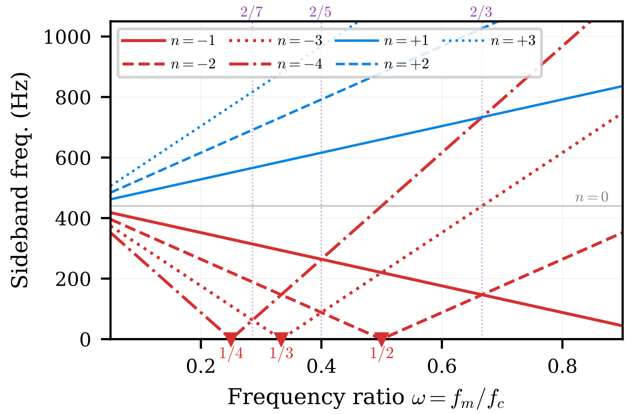


Figure 4: Folded sideband frequencies $\tilde{f}_n(\omega) = f_c|1+n\omega|$. Colored V-shaped curves show the folded sidebands, with different line styles distinguishing sideband indices and cusps at $\omega = 1/|n|$; the purple vertical dotted lines at $\omega = 2/7, 2/5, 2/3$ mark ordering transitions where two folded sidebands coincide and exchange order.

ridges ($\omega = 1, 1/2$) are prominently visible and the landscape is nearly smooth. As β increases, higher-order sidebands become significant: at $\beta = 3$, sidebands up to $|n| \approx 4$ contribute, activating ridges at $\omega = 1/3, 1/4, 2/5, 2/7$; at $\beta = 5$, sidebands up to $|n| \approx 6$ are non-negligible, further densifying the ridge pattern. This densification is asymmetric across the ω axis: in the high-ω region ($\omega \gtrsim 1$), consecutive low-order boundaries remain widely spaced regardless of β, whereas in the low-ω region ($\omega \lesssim 1/2$), newly activated higher-order rationals are inserted between already dense boundaries, further narrowing the smooth wedge regions. This β-dependent asymmetry explains the strong interaction between β and ω observed in Section 6.3. Crucially, since reducing β can only deactivate ridges, never create new ones, the full set of candidate boundaries can be enumerated a priori by computing ridge prominence at the maximum expected β.

5. METHOD

5.1. Frequency-Ratio Constraints

As shown in Section 4, the Wasserstein-1 loss landscape exhibits non-differentiable ridges at rational frequency ratios $\omega = m/n$. These ridges partition the parameter space into disjoint regions where the loss function is smooth.

To exploit this structure, we constrain the frequency ratio $\omega = f_m/f_c$ within the interior of a single smooth region during optimization. Specifically, we impose the constraint:

$$\omega \in (\omega_\ell, \omega_u) \tag{7}$$

where $\omega_\ell < \omega_u$ are consecutive low-order rational boundaries.

To enforce this constraint differentiably, we reparameterize ω using a sigmoid transformation:

$$\omega = \omega_\ell + (\omega_u - \omega_\ell) \cdot \sigma(a), \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function. Rather than optimizing ω directly, we optimize the unconstrained parameter $a \in \mathbb{R}$. The sig-

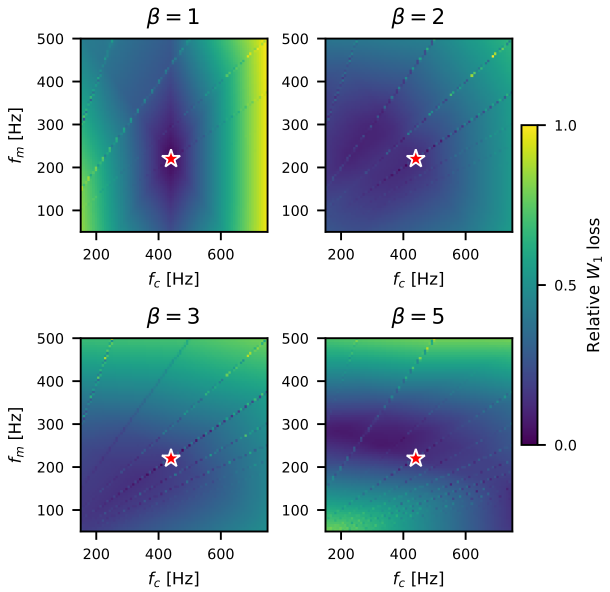


Figure 5: W_1 loss landscapes for $\beta = 1, 2, 3, 5$ ($f_c^* = 440$ Hz, $f_m^* = 220$ Hz). Panels are independently normalized for visibility. Ridge positions are β-independent; higher β activates additional ridges via higher-order sidebands.

moid maps any real value a to the open interval (ω_ℓ, ω_u), allowing standard gradient-based optimizers to operate without explicit projection steps. The modulator frequency is then computed as $f_m = \omega \cdot f_c$, ensuring that the frequency ratio constraint is satisfied throughout optimization.

By restricting ω to a single wedge-shaped region between consecutive low-order ridges, the optimizer is confined to that region while the competing local minima on all other ridges and in the other wedge interiors are structurally excluded from the search space, leaving far fewer minima for gradient-based methods to contend with.

5.2. Multi-Start Strategy

Since the true frequency ratio is unknown a priori, we run independent optimizations for each wedge region $(\omega_\ell^{(k)}, \omega_u^{(k)})$, $k = 1, \ldots, K$, each initialized at its midpoint. The solution with the lowest W_1 loss across all regions is selected as the final estimate. For target ratios lying in the interior of one wedge, this multi-start approach ensures that one subproblem contains the true ω, while the per-region cost remains low due to the constrained search space.

6. EXPERIMENTS

6.1. Experimental Setup

We synthesized FM signals at a sampling rate of $f_s = 44\,100$ Hz with $N = 65536$ samples (approximately 1.5 s). The FFT was computed with $N_{\mathrm{FFT}} = 131072$ ($2\times$ zero padding), yielding a frequency bin spacing of approximately 0.34 Hz and smoother gradients for optimization. A Hann window was applied to the signal before computing the FFT. Spectra were computed as normalized power spectra (squared FFT magnitudes).

The parameter search space was defined as $f_c \in [30, 2000]$ Hz and $f_m \in [20, 1500]$ Hz. We evaluate modulation indices $\beta \in \{1, 2, 3\}$, covering the range from near-sinusoidal to moderately complex spectra with up to 8 significant sidebands. In this range, the significant sidebands remain below Nyquist ($f_c + (\beta+1) f_m \lesssim 8$ kHz $\ll f_s/2 = 22\,050$ Hz), so the experiments isolate 0-Hz folding rather than sampled-band aliasing. Ridge boundaries were set at $\omega \in \{1/4, 2/7, 1/3, 2/5, 1/2, 2/3, 1, 2\}$, partitioning the frequency-ratio space into $K = 9$ disjoint smooth regions (including the regions below $1/4$ and above 2). These boundaries correspond to the prominent folding cusps and ordering transitions visible in Fig. 4, which arise from sidebands up to order $|n| \approx \beta + 1 = 4$ (the maximum β used in our experiments). Note that $\omega = 1$ is a ridge boundary where $f_c = f_m$, corresponding to a degenerate case that we exclude from evaluation.

All methods were implemented in PyTorch (v2.5.1), using automatic differentiation to obtain gradients of the W_1 loss with respect to the synthesis parameters. For each test case, the carrier frequency f_c and the frequency ratio ω are optimized together as free variables of a single Adam optimization (learning rate $\eta = 0.02$, 1500 iterations) over the shared W_1 loss. Since $f_m = \omega f_c$, both frequencies are thus estimated concurrently rather than in separate stages. The $K = 9$ candidates of a case are optimized in parallel as a single batched tensor operation, with each candidate updated independently. For the Constrained strategy, the sigmoid reparameterization of Eq. (8) keeps ω within its assigned wedge at every iteration, so the constraint is enforced by construction rather than by penalty or projection. On a single NVIDIA RTX A6000 GPU, one case (nine candidates, 1500 iterations) completes in approximately 3.4 s.

We compare three optimization strategies:

- **Constrained:** The frequency ratio ω is constrained to satisfy Eq. (7) throughout optimization.
- **Init-only:** Parameters are initialized within a constraint region, but no constraints are enforced during optimization.
- **Random:** Random initialization without any constraints (baseline).

For Constrained and Init-only, we place one initial point per wedge: ω starts at the wedge midpoint, and f_c at the midpoint of the feasible carrier range satisfying $f_m = \omega f_c \in [f_{m,\min}, f_{m,\max}]$. This placement avoids the ridge boundaries where discontinuities occur and ensures that both constrained methods start from equivalent positions, isolating the effect of maintaining constraints during optimization. For the Random baseline, the same number of initial points ($K = 9$) are drawn uniformly from the full (f_c, f_m) search box.

6.2. Evaluation

We generated 90 test cases using a factorial design with three factors: carrier frequency $f_c \in \{200,\ 400,\ 700\}$ Hz, frequency ratio $\omega \in \{1/4,\ 0.30,\ 1/3,\ 2/5,\ 1/2,\ 0.55,\ 2/3,\ 0.85,\ 1.5,\ 1.8\}$, and modulation index $\beta \in \{1, 2, 3\}$. The frequency-ratio grid includes exact ridge boundaries ($\omega = 1/4, 1/3, 2/5, 1/2, 2/3$) and nearby off-boundary values (0.30, 0.55), intentionally stress-testing boundary cases. Each test case was evaluated over three independent trials for the Random baseline. The estimation quality for each parameter is measured as the relative error in percent, e.g., $100 \cdot |\hat{\omega} - \omega^*|/\omega^*$ for the frequency ratio.

Table 1: Parameter estimation errors (%) and success rate. Errors show median / p90. Succ. = fraction with ω error $< 1\%$. Best in bold.

Method	Succ. ↑	ω ↓	f_c ↓	f_m ↓
Constrained	**96.7%**	**0.11 / 0.64**	0.10 / **0.42**	0.16 / **0.56**
Init-only	80.0%	0.22 / 11.38	**0.08** / 7.48	**0.14** / 14.46
Random	68.1%	0.25 / 44.99	0.09 / 24.59	0.17 / 24.95

Table 1 summarizes the results. We define success as achieving ω error below 1%. The Constrained method succeeds in 96.7% of trials, compared with 80.0% for Init-only and 68.1% for Random, demonstrating substantially higher reliability. Notably, the Constrained method exhibits a strongly bimodal error distribution: all trials either converge to sub-1% error or fail with error exceeding 10%, with no intermediate cases. As discussed further in Section 6.3, the sharp failure cluster is concentrated in the intentionally included worst-case low-order rational boundary cases, while boundary and near-boundary cases more generally account for the largest residual errors. This sharp separation makes the success rate insensitive to the exact threshold choice.

The three methods differ not in typical accuracy but in failure severity. The median errors are comparable across all methods, indicating that the majority of test cases converge successfully regardless of strategy. However, the distributions are strongly right-skewed (skewness>4), shown in Fig. 6: while the interquartile ranges for Constrained and Init-only are near zero, the upper tails diverge sharply. The 90th percentile (p90) captures this divergence: the Constrained method maintains sub-1% p90 errors across all parameters, whereas Init-only and Random exhibit p90 errors up to 14% and 45%, respectively. In practical terms, the p90 reflects how severe a one-in-ten worst case is, a quantity directly relevant to the reliability of an automatic parameter estimation system. The similarity in medians partly reflects the dominance of easy cases (low β); Section 6.3 examines how performance varies with β and ω.

Table 2 reports magnitude-based audio-domain metrics: Spectral Optimal Transport loss (SOT, i.e., W_1 between normalized power spectra), log-Mel spectrogram MSE (Mel), and Log Spectral Distance (LSD) in dB. We follow the metric set of Peladeau et al. [9], but the values are not directly comparable because their setting fixes the carrier frequency and estimates the modulation index and frequency ratio, whereas ours fixes β and jointly estimates f_c and f_m. We omit SI-SDR because even sub-percent carrier-frequency errors cause accumulated phase drift, making time-domain scores unreliable in our setting.

The median values are comparable across all three methods; the Constrained method shows slightly higher median SOT than Init-only and Random, likely because the sigmoid reparameterization introduces minor residual error when the true ω lies on or near a wedge boundary. However, the p90 values reveal the same pattern as Table 1: the Constrained method maintains substantially lower worst-case spectral errors across all three metrics (Table 2), confirming that the constraint improves reliability at the cost of negligible typical-case overhead.

6.3. Discussion

Figure 7 illustrates representative optimization trajectories for a challenging case ($f_c = 400$ Hz, $\omega = 0.25$, $\beta = 3$). Each panel

Table 2: Audio-domain spectral metrics. SOT: Spectral Optimal Transport (W_1); Mel: log-Mel MSE; LSD: Log Spectral Distance (dB). Best in **bold**.

	SOT ↓		Mel ↓		LSD ↓	
Method	med.	p90	med.	p90	med.	p90
Constrained	0.76	**29.61**	**0.42**	**5.60**	**3.73**	**7.80**
Init-only	0.72	38.69	0.67	34.62	3.89	9.24
Random	**0.69**	56.83	0.52	98.87	4.28	11.96

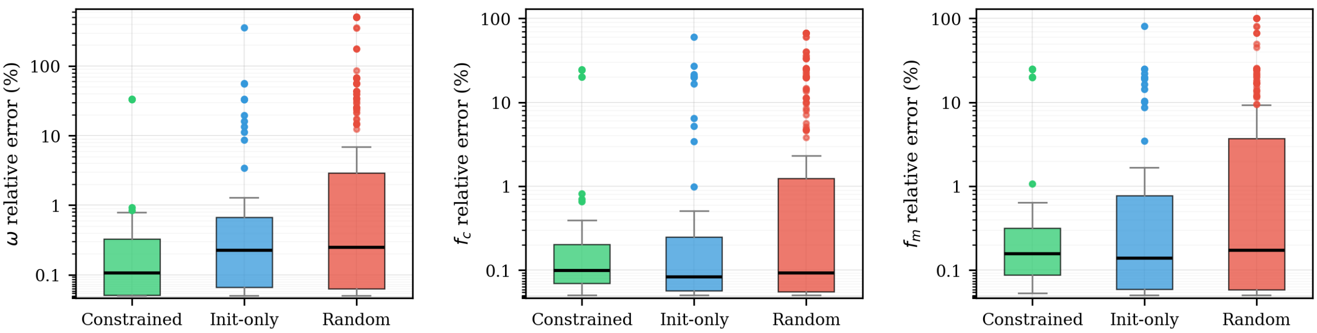
shows the W_1 loss landscape in the (f_c, f_m) plane with the ground truth marked by a green circle. Yellow stars indicate initial points, colored curves show optimization paths, and the red star marks the selected best solution. White dashed lines denote the rational ridge boundaries $\omega = m/n$.

The Constrained method (left) successfully converges to the ground truth: each trajectory remains within its designated wedge region, avoiding ridge crossings. In contrast, the Init-only method (center) starts from the same well-placed initial points but crosses ridge boundaries during optimization, becoming trapped in a local minimum in an adjacent wedge. The Random method (right) initializes at arbitrary locations, traverses multiple ridges, and becomes trapped in a smooth interior basin far from the ground truth, illustrating that traps arise not only at ridges but also within wedge interiors.

This visual comparison highlights the core insight: maintaining the frequency-ratio constraint throughout optimization, not just at initialization, is essential for reliable convergence in complex loss landscapes.

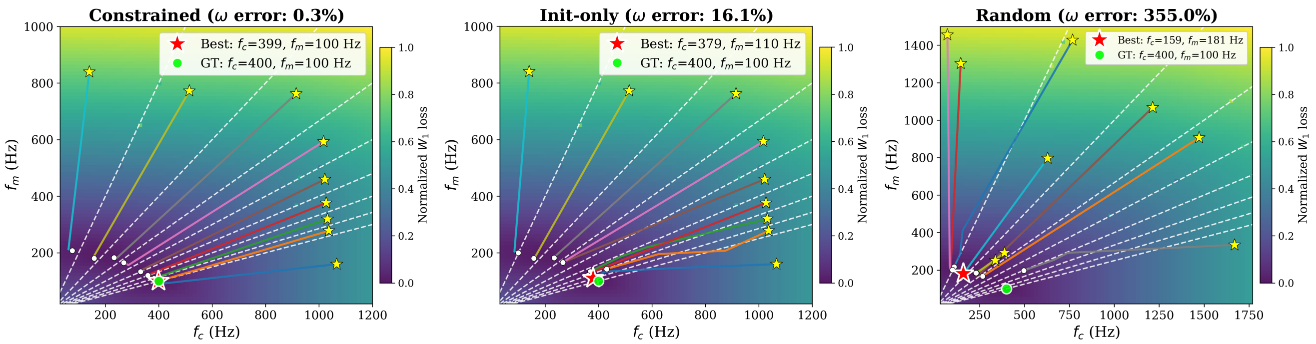
The experimental results reveal a strong interaction between β and the effectiveness of each method (Fig. 8). At $\beta = 1$, all three methods achieve comparable error distributions, whereas at $\beta = 3$, the Constrained method substantially outperforms Random. Quantitatively, the p90 ω errors at $\beta = 3$ are 4.2%, 35.5%, and 351.4% for Constrained, Init-only, and Random, respectively, compared with 0.3%, 1.1%, and 2.0% at $\beta = 1$. As analyzed in Section 4, higher β activates more sidebands and densifies the ridge pattern, increasing the density of spurious local minima that trap unconstrained optimizers. The Constrained method is robust to this effect because each subproblem is confined to a single wedge, excluding the spurious minima associated with all other ridges regardless of spectral complexity.

The interaction with frequency ratio ω has a geometric explanation (Fig. 9). The box plots confirm that errors are concentrated in the low-ω region ($\omega < 0.5$), where Random and Init-only exhibit wide distributions with large outliers, while the Constrained method maintains tight distributions across all bands. For $\omega < 0.5$, the p90 ω errors are 0.7%, 33.3%, and 177.8% for Constrained, Init-only, and Random; for $\omega \geq 1$, the corresponding p90 errors are only 0.1%, 0.3%, and 0.1%. Since ridges in the (f_c, f_m) plane are radial lines through the origin (Fig. 2), the distance between consecutive ridges at a given f_c is $f_c \cdot \Delta\omega$. At low ω, where rational boundaries are densely spaced, the wedge regions become very narrow: for example, at $f_c = 400$ Hz, the wedge between $\omega = 1/4$ and $\omega = 2/7$ spans only $400 \times 1/28 \approx 14$ Hz in f_m, whereas the wedge between $\omega = 1$ and $\omega = 2$ spans 400 Hz. These narrower wedge regions make it easier for unconstrained optimization trajectories to cross ridge boundaries and become trapped in adjacent wedges, disproportionately affecting the Ran-



(a) ω relative error (%) **(b)** f_c relative error (%) **(c)** f_m relative error (%)

Figure 6: Parameter estimation relative errors (%) on a log scale. Box plots show median (line), interquartile range (box), 1.5×IQR whiskers, and individual outliers. The interquartile ranges for Constrained and Init-only are near zero; the methods differ primarily in the severity of the upper tail, which is captured by the p90 values in Table 1.



(a) Constrained **(b)** Init-only **(c)** Random

Figure 7: Optimization trajectories for $f_c = 400$ Hz, $\omega = 0.25$, $\beta = 3$. Green circle: ground truth; yellow stars: initial points; red star: best solution; trajectories connect sampled steps at 300, 800, and 1500 iterations. Panels (a) and (b) use a reduced plotting range, while panel (c) retains the full range.

dom and Init-only methods. They also reduce the admissible range of ω itself, so the constrained subproblem approaches an effectively one-dimensional search in f_c within a fixed local ratio band. The Constrained method is also affected to a lesser extent: in very narrow wedges, the sigmoid reparameterization maps to a small interval, which can cause gradient attenuation near the boundaries. Indeed, all failures of the Constrained method (9 out of 270 trials) occur at $\omega = 1/3$ with $\beta = 3$, where the true ω lies on a ridge boundary and therefore on the edge of two adjacent wedges. More broadly, the largest residual errors are concentrated on or near the rational-ratio boundaries that were intentionally included as worst-case stress-test conditions. In these cases, neither wedge contains the true ω in its interior, and the multi-start selection instead converges to a competing low-ratio local minimum ($\hat{\omega} \approx 0.22$). This limitation arises from the current non-overlapping boundary-based partitioning strategy and could be mitigated by including overlapping wedge regions around each boundary.

As noted in Section 4, the ridge positions are β-independent, so the wedge boundaries used in our method do not require precise knowledge of β. Since ridge prominence increases monotonically with β for each boundary (higher β activates more sidebands), computing boundaries at the maximum expected β yields a conservative partition valid for all lower values, at the cost of slightly more wedge subproblems.

7. CONCLUSION AND FUTURE WORK

We presented a gradient-based optimization strategy for FM parameter estimation that exploits the effectively smooth structure of the Wasserstein loss landscape between consecutive low-order rational boundaries. By analyzing the loss landscape of 2-operator FM synthesis under FFT-based spectral representations, we identified non-differentiable ridges at rational frequency ratios $\omega = m/n$ arising from frequency folding and ordering transitions. Our proposed method constrains the frequency ratio within intervals bounded by consecutive low-order rational boundaries, improving convergence reliability in the tested setting. Experimental results from controlled ablations showed that maintaining the constraint throughout optimization improves reliability over random initialization and initialization-only strategies, particularly for more complex spectra at higher modulation indices.

Preliminary investigation of 3-operator FM synthesis suggests that the ridge structure in its loss landscape can be analytically

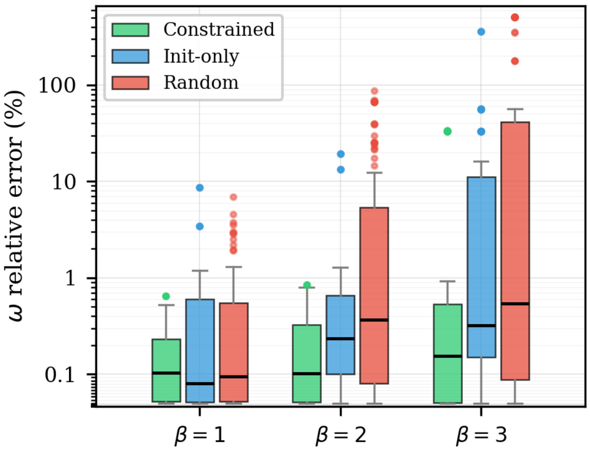


Figure 8: Distribution of ω relative error (%) by modulation index β (log scale). Box plots show median, IQR, and outliers.

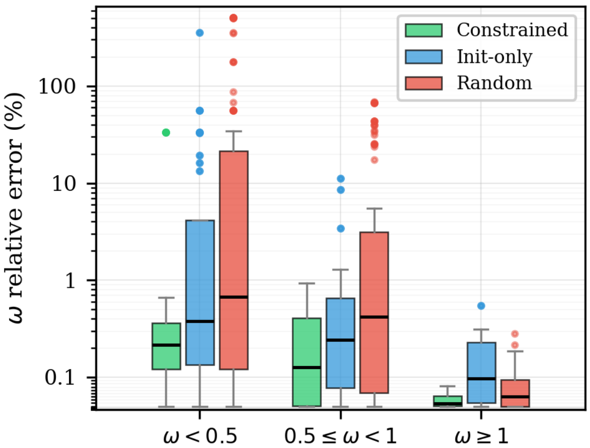


Figure 9: Distribution of ω relative error (%) by frequency ratio band (log scale). Box plots show median, IQR, and outliers.

characterized in a similar manner to the 2-operator case. For example, a 2-dimensional cross-section of the loss landscape over the two modulator frequency ratios $(f_{m_1}/f_c,\ f_{m_2}/f_c)$ reveals that, in addition to rational ratios of each modulator frequency to the carrier, ridges also arise at rational combinations of the two modulator frequencies. Future work includes higher-order FM and joint estimation of modulation indices and envelopes. Another direction is to analyze aliasing regimes where significant sidebands cross Nyquist and may introduce additional ridges. Since the present study used FM-synthesized targets only, future work should also address arbitrary audio targets and perceptual listening evaluations.

8. ACKNOWLEDGMENTS

The authors thank Masahiro Nakanishi and Maki Kato for helpful discussions.

9. REFERENCES

[1] J. M. Chowning, “The synthesis of complex audio spectra by means of frequency modulation,” *J. Audio Eng. Soc.*, vol. 21, no. 7, pp. 526–534, Sep. 1973.

[2] J. O. Smith III, *Spectral Audio Signal Processing*. W3K Publishing, 2011, [Online]. Available: https://ccrma.stanford.edu/~jos/sasp/ (accessed: Mar. 2026).

[3] A. Horner, J. Beauchamp, and L. Haken, “Machine tongues XVI: Genetic algorithms and their application to FM matching synthesis,” *Comput. Music J.*, vol. 17, no. 4, pp. 17–29, 1993.

[4] T. J. Mitchell and J. C. W. Sullivan, “Frequency modulation tone matching using a fuzzy clustering evolution strategy,” in *Audio Engineering Society 118th Convention*, Barcelona, Spain, May 2005, convention Paper 6366.

[5] Q. Legros, D. Fourer, S. Meignen, and M. A. Colominas, “Instantaneous frequency and amplitude estimation in multicomponent signals using an EM-based algorithm,” *IEEE Trans. Signal Process.*, vol. 72, pp. 1130–1140, 2024.

[6] M. Caetano, “Parameter estimation of frequency-modulated sinusoids with the distribution derivative method,” in *Proc. Int. Conf. Digital Audio Effects (DAFx24)*, Guildford, Surrey, UK, 2024, pp. 322–328.

[7] Z. Chen, Y. Jing, S. Yuan, Y. Xu, J. Wu, and H. Zhao, “Sound2Synth: Interpreting sound via FM synthesizer parameters estimation,” in *Proceedings of the 31st Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2022, pp. 4921–4928.

[8] B. Hayes, C. Saitis, and G. Fazekas, “Audio synthesizer inversion in symmetric parameter spaces with approximately equivariant flow matching,” in *Proceedings of the 26th Int. Society for Music Information Retrieval Conf. (ISMIR)*, Daejeon, South Korea, 2025.

[9] C. Peladeau, D. Fourer, and G. Peeters, “Audio processor parameters: Estimating distributions instead of deterministic values,” in *Proc. Int. Conf. Digital Audio Effects (DAFx25)*, Ancona, Italy, 2025, pp. 275–282.

[10] F. Caspe, A. McPherson, and M. Sandler, “DDX7: Differentiable FM synthesis of musical instrument sounds,” in *Proceedings of the 23rd Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2022, pp. 608–616.

[11] J. Turian and M. Henry, “I’m sorry for your loss: Spectrally-based audio distances are bad at pitch,” in *NeurIPS 2020 ICBINB Workshop*, 2020.

[12] B. Hayes, C. Saitis, and G. Fazekas, “Sinusoidal frequency estimation by gradient descent,” in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] B. Torres, G. Peeters, and G. Richard, “Unsupervised harmonic parameter estimation using differentiable DSP and spectral optimal transport,” in *IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1176–1180.

[14] N. Matsubayashi, “Wasserstein inverted frequency modulation synthesizer,” Presentation slides, Speaker Deck, Mar. 2021, presented at Kernel/VM Exploration Team Online Part 2. (in Japanese) [Online]. Available: https://speakerdeck.com/fadis/wassersteinni-fmyin-yuan (accessed: May 2026).

[15] C. Villani, *Optimal Transport: Old and New*. Berlin, Heidelberg: Springer, 2009.

[16] F. Santambrogio, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Cham: Birkhäuser, 2015.