

Discrete Parameter Estimation for DDSP Source Separation

David Millard*, Andrea Cogliati



University of Rochester

Introduction

Differentiable digital signal processing (DDSP) fits the parameters of a physical synthesizer by gradient descent on a spectral loss. That works well for continuous controls. Although, many physical models, are controlled by integer delays. To parameterize an integer requires discretization, which has a gradient zero everywhere. Without an approximation/relaxation we are unable to parameterization the exponentials.

We study this in its simplest form: a Karplus–Strong string with continuous loop gain K and all-pass coefficient α , plus two integer parameters:

- Loop delay L , which together with α sets the pitch: $f_0 \approx \frac{f_s}{L+1+\alpha}$
- Pluck position, defined in two ways: a comb-filter delay p (under excitation $-K(\delta[n] - \delta[n-p])$), or the apex A of a triangular initial string shape

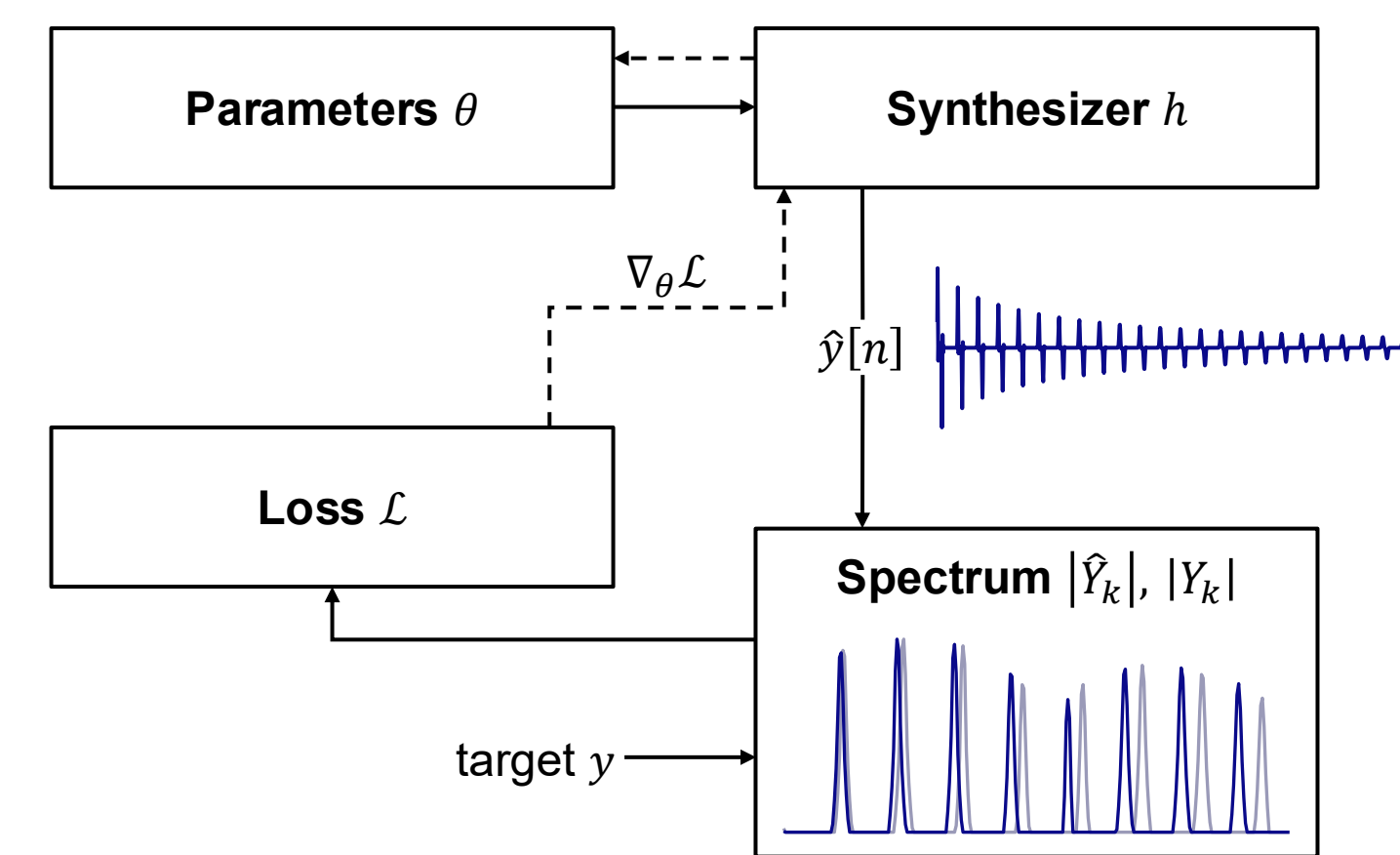
We compare classical estimators (exhaustive search, pitch estimation, continuous relaxation) with learned categorical estimators (REINFORCE, Gumbel-Softmax). We evaluate mixtures of up to 7 strings, where the task becomes source separation and multi pitch estimation.

What is DDSP?

Differentiable digital signal processing (Engel et al., 2020) builds a synthesizer $\hat{y} = h(\theta)$ from classic blocks such as oscillators, filters, and delays, such that every operation is differentiable. Its parameters θ are fit to a recording y by comparing spectra,

$$\mathcal{L}(\theta) = \sum_k |\log|\hat{Y}_k| - \log|Y_k||,$$

and performing gradient descent to carry error through the synthesizer.

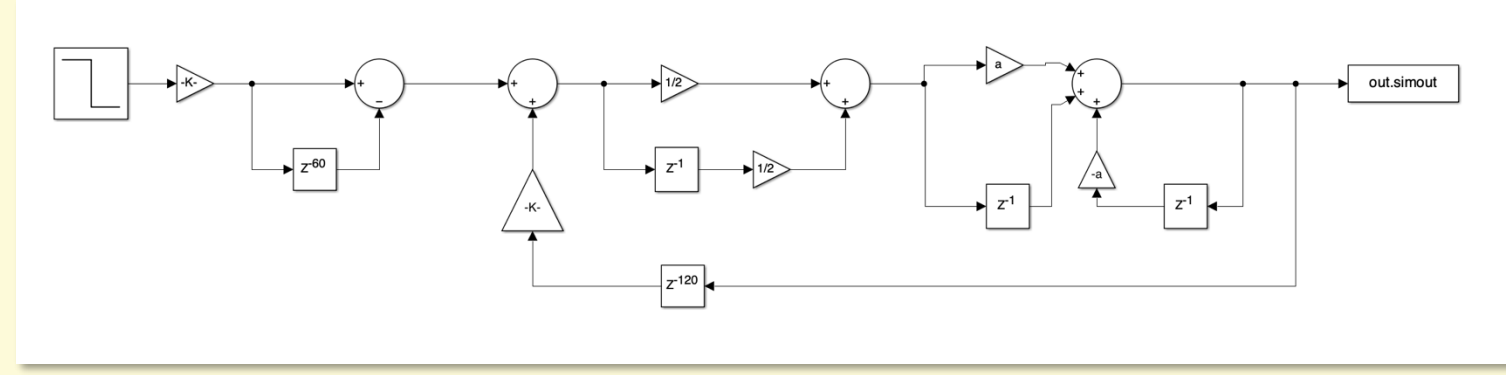


Our synthesizer is the extended Karplus–Strong model,

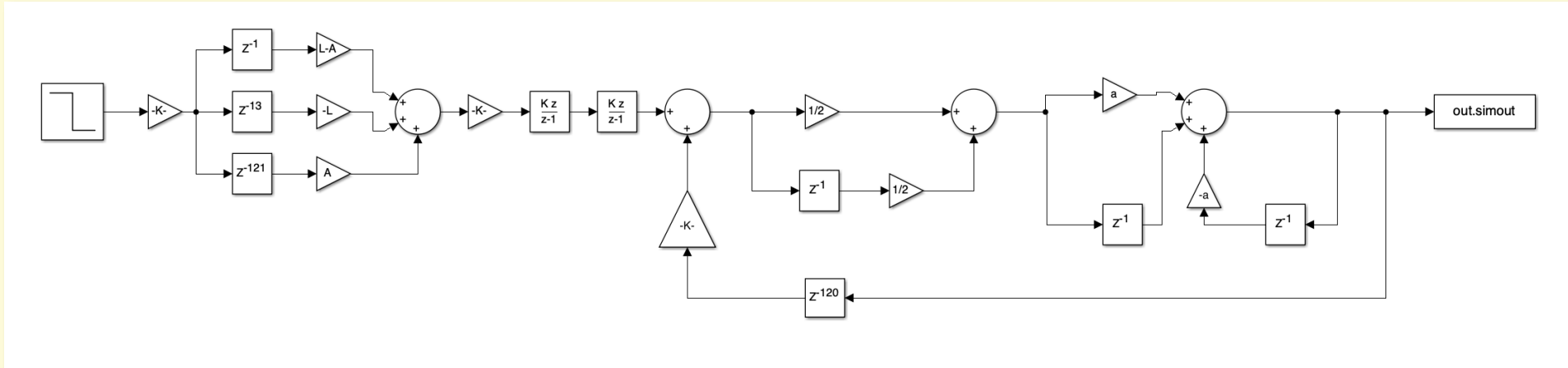
$$E(z; p) H(z) = E(z; p) \frac{(1 + z^{-1})(a + z^{-1})}{2(1 + \alpha z^{-1}) - K z^{-L}(1 + z^{-1})(a + z^{-1})}.$$

where $E(z; p)$ is a filter controlling pluck position. The gain K and all-pass coefficient α are continuous, so backpropagation works great! But since the delay length L and pluck position p are integers, neither $\partial \hat{y} / \partial L$ or $\partial \hat{y} / \partial p$ are defined, due to their gradient being zero almost everywhere.

Models

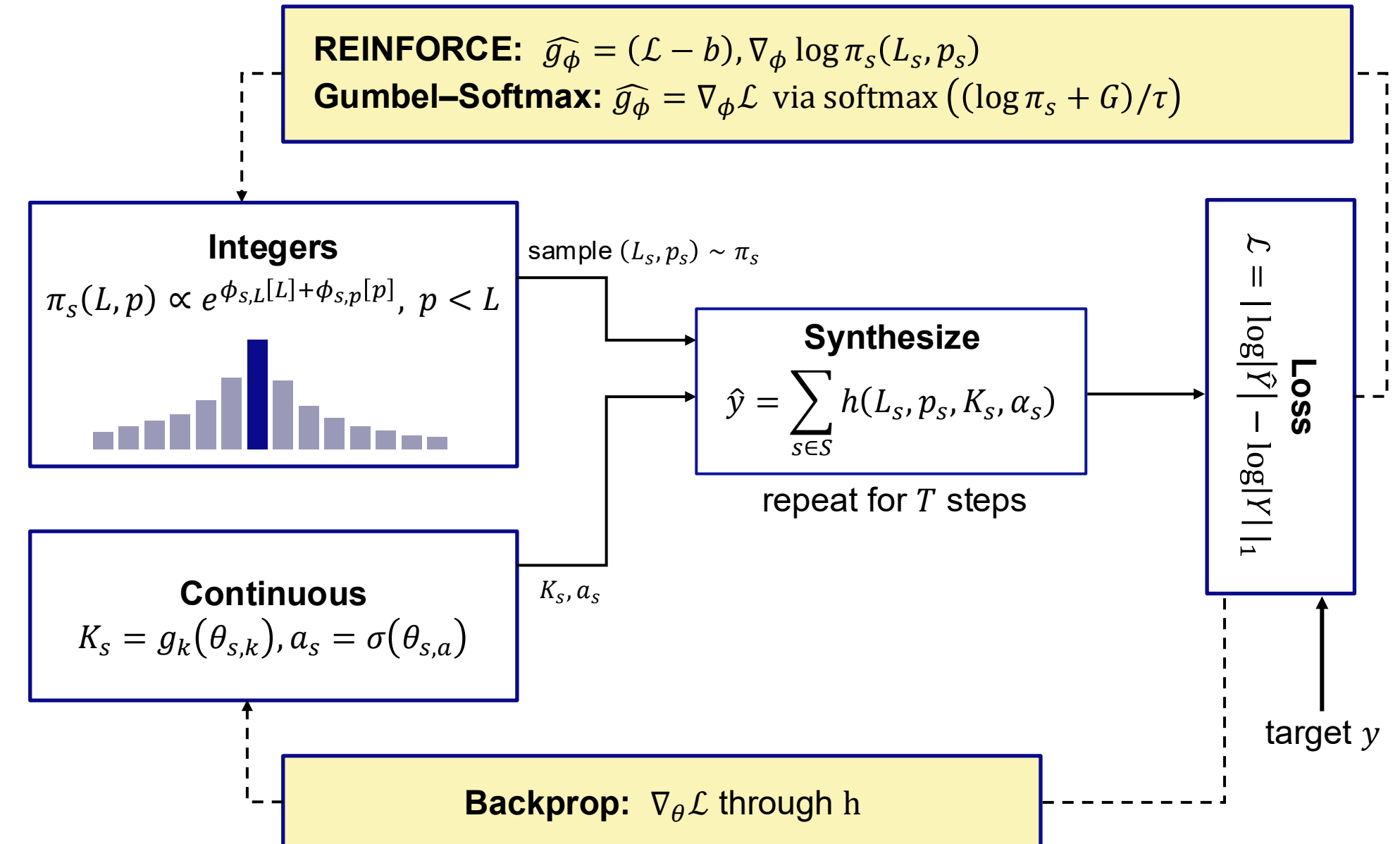


Model 1. A feedforward comb on the excitation sets the pluck position and notches out the harmonics that have a node at that point.



Model 2. Weighted delays followed by double integration build the plucked string's triangular starting shape, with its peak at the pluck position.

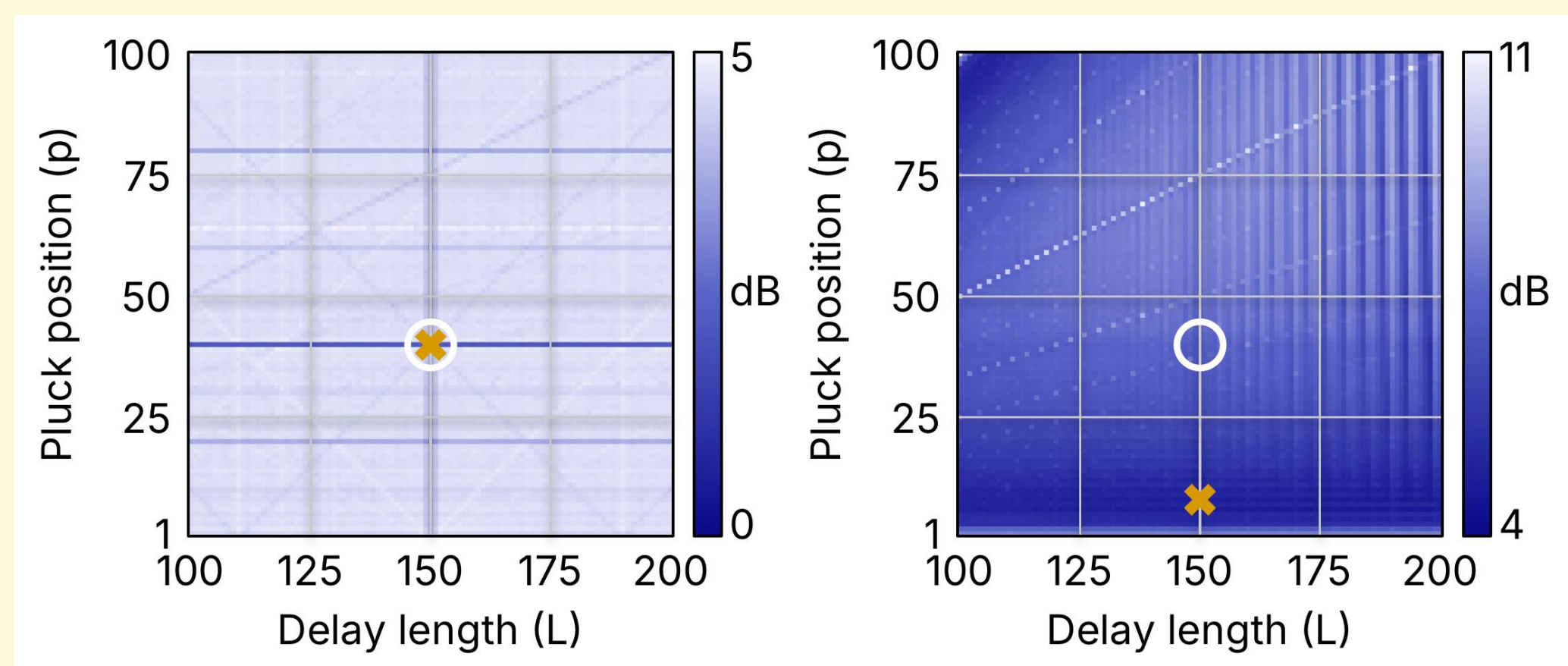
Algorithm



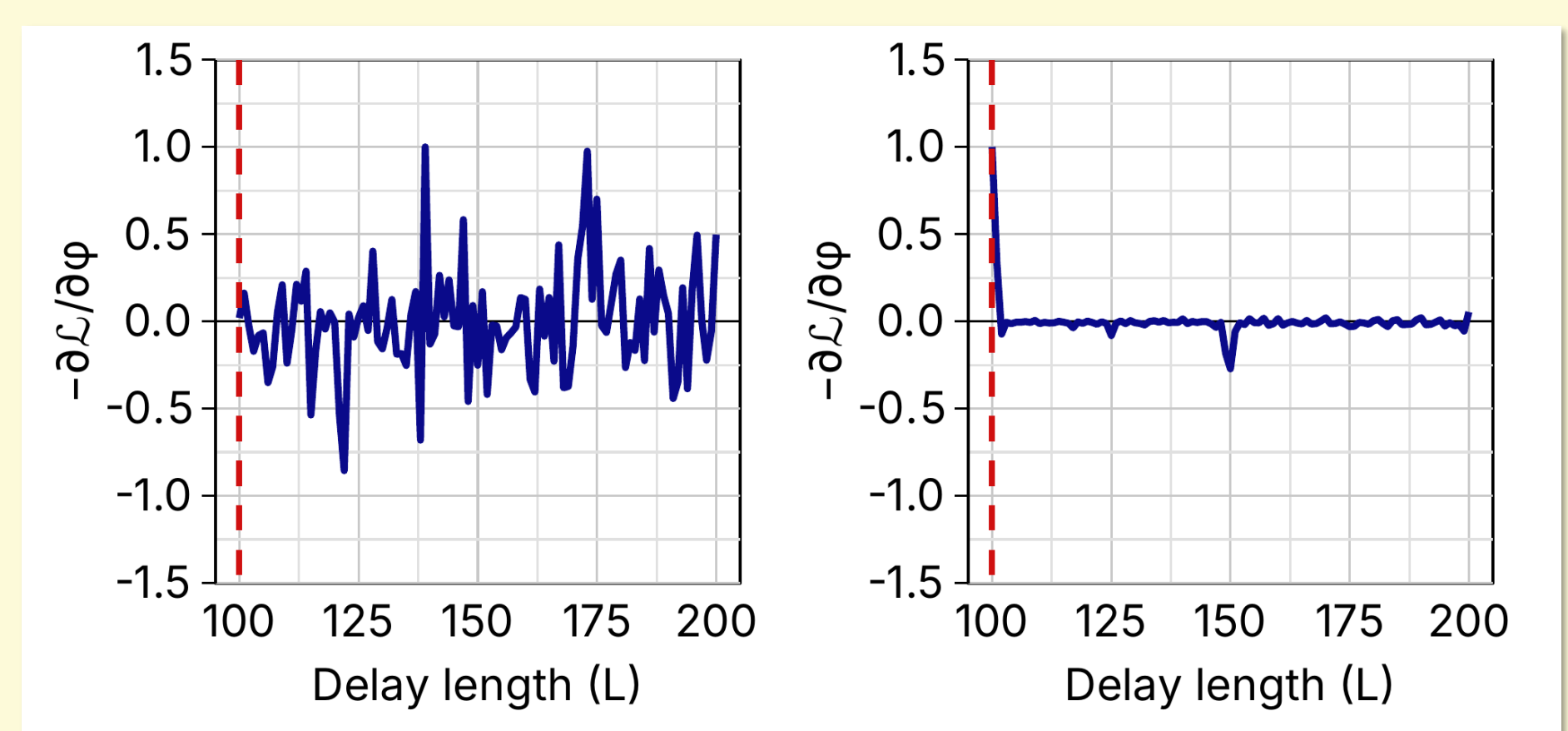
Results

Comb excitation.														Triangle excitation.														
exact (L, p), %														pitch F1, %														
Reinforce	79	93	95	87	85	86	84	100	100	100	100	98	98	97	66	46	43	35	40	32	31	92	84	74	76	76	78	72
Gumbel-sampled	93	95	98	92	86	69	52*	100	100	100	100	98	82	70*	70	54	42	22	18	12*	7	95	87	71	61	52	50*	43
Gumbel	0	0	0	0	1	1	0	8	16	20	33	42	41	43	0	2	0	2	0	0	0	23	41	29	30	29	31	29
Relaxation	0	0	1	1	0	1	1	8	21	33	44	43	45	40	0	0	1	0	1	1	1	21	32	32	36	39	42	43
Lagrange Fractional Delay	0*	0*	0*	0*	0*	0*	0*	0*	10*	7*	5*	12*	10*	0*	0*	0*	0*	0*	0*	0*	0*	20*	0*	0*	20*	12*	3*	11*
Round + STE	0*	0*	0*	0*	0*	0*	0*	0*	10*	13*	0*	12*	17*	3*	0*	0*	0*	0*	0*	0*	0*	20*	10*	0*	15*	16*	3*	9*
Exhaustive Search	100	—	—	—	—	—	—	100	—	—	—	—	—	—	53	—	—	—	—	—	—	97	—	—	—	—	—	—
Klapuri	pitch only							100	75	70	61	55	43	45	pitch only							100	74	60	54	48	46	39
Autocorrelation	pitch only							97	97	94	87	83	81	74	pitch only							100	43	42	27	23	27	21
Random Guess	pitch only							8	15	21	27	32	35	37	pitch only							9	15	21	26	31	35	38
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7

Method comparison on comb (left) and triangle (right) mixtures of 1–7 sources. Cells give the percentage of sources with the exact (L, p) recovered and the pitch F1 within 50 cents. Each cell averages 32 recordings $\times$ 5 seeds; * marks cells with fewer runs.



Spectral loss with misaligned (L, p, K) pairs for comb (left) and triangle (right) excitation, using a gain of $K = 0.9$ when the true gain is 0.95. For comb, the lowest loss (X) stays at the true pair (O). For triangle, it moves to the wrong pluck position.



Updates to the delay logits ϕ from original (left) and sampled (right) Gumbel-Softmax on one comb recording, scaled to a peak of 1; red line marks the true delay. Only the sampled update favors the true delay.

What makes discrete values difficult to optimize?

Gradient descent breaks down on discrete values because quantizing a value destroys its gradient. Consider the simplest quantizer, rounding a real number x to the nearest integer:

$$r(x) = \lfloor x + 1/2 \rfloor$$

where $\lfloor \cdot \rfloor$ rounds down. The output $r(x)$ is flat between half-integers and jumps by 1 at each half-integer, **Figure 1**. Its slope $\partial r / \partial x$ is therefore 0 everywhere except at the jumps, where it does not exist, **Figure 2**.

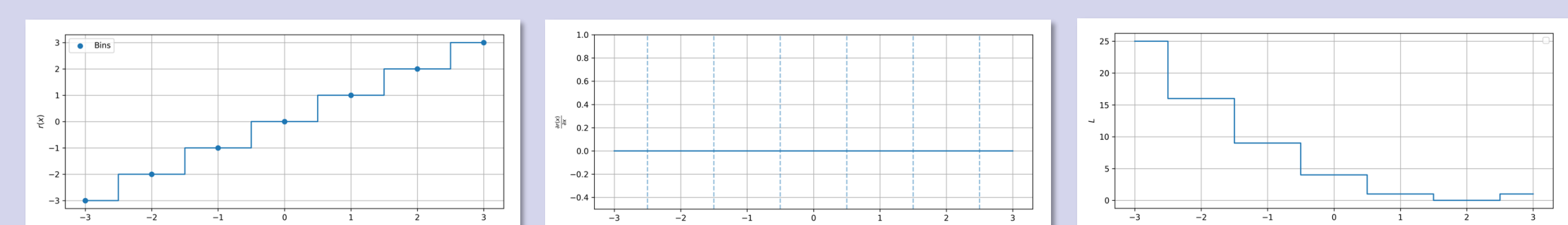


Figure 1. $r(x)$ shown as a function of x .

Figure 2. $\frac{\partial r(x)}{\partial x}$ shown as a function of x .

Figure 3. $\mathcal{L}$ shown as a function of x .

Forward pass. Suppose we want the rounded value to equal 2. We measure the error with the squared loss,

$$\mathcal{L}(x) = (r(x) - 2)^2.$$

Backward pass. The chain rule splits the gradient of $\mathcal{L}$ into the slope of the loss and the slope of the rounding:

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial r} \cdot \frac{\partial r}{\partial x}, \quad \frac{\partial \mathcal{L}}{\partial r} = 2(r(x) - 2).$$

Non-gradient Baselines

Exhaustive search. Synthesize every valid pair and keep the one with the lowest loss:

$$(\tilde{L}, \tilde{p}) = \operatorname{argmin}_{(L, p) \in \mathcal{G}} \mathcal{L}(L, p; K_0, a_0)$$

where $\mathcal{G}$ is the set of valid pairs and K_0, a_0 are fixed starting values.

Strength: exact for one source. **Weakness:** scales exponentially.

Pitch estimation. Estimate each source's fundamental frequency $\hat{f}_{0,s}$ from the mixture, then convert its period to a loop length:

$$\tilde{L}_s = \operatorname{round}\left(\frac{f_{\text{SR}}}{\hat{f}_{0,s}} - \frac{1}{2} - \frac{1 - a_s}{1 + a_s}\right)$$

where f_{SR} is the sample rate, and the last two terms remove the delay of the loop filter.

Strength: fast, efficient, and established. **Weakness:** gives L_s only.

Relaxed Methods

Replace each integer delay by a real number $\tilde{L}_s, \tilde{p}_s$, follow the gradient $\partial \mathcal{L} / \partial \tilde{L}_s$, and round down at the end, $L_s = \lfloor \tilde{L}_s \rfloor$.

Continuous relaxation. In the frequency domain a delay is a phase shift, which is defined for any real delay:

$$z_k^{-\tilde{L}} = e^{-j\omega_k \tilde{L}}$$

where ω_k is the frequency of DFT bin k .

Lagrange fractional delay. A 6-tap FIR filter interpolates between neighboring integer delays:

$$z^{-\tilde{L}} \approx \sum_{k=0}^5 h_k(\delta) z^{-(n+k)}, \quad h_k(\delta) = \prod_{j \neq k} \frac{\delta - j}{k - j}$$

where n is an integer just below $\tilde{L}$ and $\delta = \tilde{L} - n$.

Round + straight-through. Round in the forward pass, but treat rounding as the identity in the backward pass:

$$L_s = \operatorname{round}(\tilde{L}_s), \quad \frac{\partial L_s}{\partial \tilde{L}_s} = 1.$$

Categorical Estimators

Each source s puts a probability on every valid pair $j = (L, p)$ and minimizes the expected loss J :

$$\pi_s(j) \propto e^{\phi_j}, \quad \phi_j = \phi_{s,L}[L] + \phi_{s,p}[p], \quad J = \mathbb{E}_{j \sim \pi_s}[\mathcal{L}_j], \quad \frac{\partial J}{\partial \phi_j} = \pi_s(j)(\mathcal{L}_j - J),$$

where $\mathcal{L}_j$ is the spectral loss of pair j . The exact gradient needs all paired losses, so each estimator substitutes something cheaper for $\mathcal{L}_j$.

REINFORCE. Sample $M = 8$ pairs from π_s , synthesize each one, and lower the probability of pairs whose loss is above average: **Sampled Gumbel.** Add Gumbel noise g to the logits, synthesize $M = 8$ pairs with the largest $\phi_j + g_j$, and weigh their measured losses:

$$\hat{g} = \frac{1}{M} \sum_{m=1}^M (\mathcal{L}_m - b_m) \nabla_{\phi} \log \pi_s(j_m), \quad \tilde{L} = \sum_{m=1}^M w_m \mathcal{L}_m, \quad \frac{\partial \tilde{L}}{\partial \phi_m} = \frac{w_m (\mathcal{L}_m - \tilde{L})}{\tau}$$

where $j_m \sim \pi_s$ are the sampled pairs, $\mathcal{L}_m$ their measured losses, and b_m the mean loss of the other 7 samples.

Strength: unbiased. **Weakness:** high variance.

where w_m is the softmax of $(\phi_m + g_m)/\tau$ over the 8 pairs, with temperature τ .

Strength: matches the exact gradient. **Weakness:** biased.

Result. The loss slope $\partial \mathcal{L} / \partial r$ points toward the target at every point while multiplying by $\partial r / \partial x = 0$ erases it. Gradient descent never moves x , even at $x_3 = 3$, where $\mathcal{L}_3 = 1$ is not the minimum.