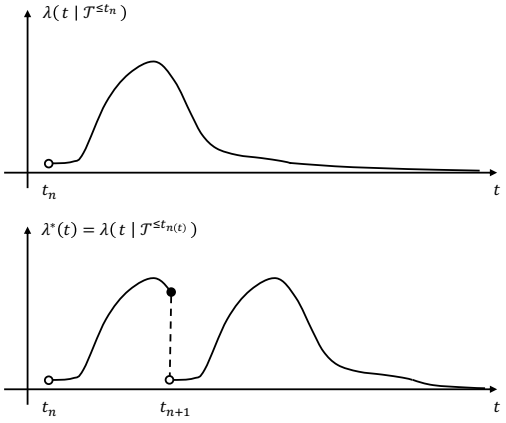

Figure 5. Conditional intensity functions of our definition (top) and the standard one (bottom). The standard conditional intensity function λ^* can be obtained by joining multiple conditional intensity functions of our definition in a left-continuous way.

A. Relation to the Standard Notation

Our notation is different from the standard one employed by many others including a textbook (Daley & Vere-Jones, 2003). A major difference is the conditional intensity function as illustrated in Figure 5. Our conditional intensity function $\lambda(t \mid \mathcal{T}^{\leq t_n})$ is defined for $t > t_n$ and is consistently conditioned on the history of events up to t_n (the top panel of Figure 5). On the other hand, the standard conditional intensity function $\lambda^*(t)$ is defined for all t, and the history that conditions it is ambiguous and depends on the context; it is sometimes conditioned on the history of events that occured before (and not including) t, which is represented as $\mathcal{T}^{\leq t_{n(t)}}$ (the bottom panel of Figure 5), and it is sometimes equivalent to our definition of the conditional intensity function.

While such an ambiguity helps to simplify equations (*e.g.*, the compensator $\Lambda^{[0,T]} = \int_0^T \lambda^*(t)\mathrm{d}t$ can be represented by a single integral), it is sometimes very confusing especially for those who are not familiar with temporal point processes. Therefore, we employ a less ambiguous definition. In order to represent the standard conditional intensity function by our conditional intensity function, we introduced the left-continuous counting process $n(t)$ as illustrated in Figure 1; with this counting process, we obtain the relationship between the standard and our conditional intensity functions, $\lambda^*(t) = \lambda(t \mid \mathcal{T}^{\leq t_{n(t)}})$.

B. Conditions for Conditional Intensity Function

This section provides a proof of Proposition 2, which states several conditions under which the conditional intensity function can specify a marked point process uniquely.

Proof of Proposition 2. Since Equation (3) reads as,

$$\begin{aligned}\int_X \mathrm{d}\mathbf{p}\lambda(t, \mathbf{p} \mid \mathcal{T}_{\bar{X}}^{\leq t_n}) &= \frac{f\left(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n}\right)}{1 - F\left(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n}\right)} \\ &= -\frac{\mathrm{d}}{\mathrm{d}t}\log(1 - F(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n})),\end{aligned}$$

we can represent the cumulative distribution function by the conditional intensity function as,

$$F(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n}) = 1 - \exp\left(-\int_{t_n}^{t} \mathrm{d}s \int_X \mathrm{d}\mathbf{p}\lambda(s, \mathbf{p} \mid \mathcal{T}_{\bar{X}}^{\leq t_n})\right). \tag{14}$$

In order for the conditional intensity function to define a proper cumulative distribution function, the function $F(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n})$ defined as Equation (14) must satisfy the following four conditions:

1. $\lim_{t\to\infty} F(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n}) = 1$,
2. $\lim_{t\to-\infty} F(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n}) = 0$,
3. $F(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n})$ is non-decreasing in t, and
4. $F(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n})$ is right-continuous.

Condition 1 is satisfied by Assumption 2. Condition 2 is satisfied because $F(t \mid \mathcal{T}_{\bar{X}}^{\leq t_n}) = 0$ holds for any $t \leq t_n$. Condition 3 holds because $\lambda(t, \mathbf{p}) \geq 0$ implies that the integral in the exponential function in Equation (14) is non-decreasing. Condition 4 holds because Assumption 3 ensures that the exponent in Equation (14) is right-continuous, which indicates that Equation (14) itself is right-continuous. □

C. Properties of Differentiable Point Processes

This section provides several properties of differentiable point processes and their proofs. First, let us prove Proposition 4, which clarifies the conditional intensity function of a differentiable point process.

Proof of Proposition 4. The following calculus clarifies the relationship:

$$\begin{aligned}&\Pr\left[t_{n+1} \in [t, t+\mathrm{d}t], \mathbf{p}_{n+1} = \mathbf{p} \mid \mathcal{T}_{\mathrm{conv}_0(\mathbb{1}^D)}^{<t}\right] \\ =&\Pr\left[t_{n+1} \in [t, t+\mathrm{d}t] \mid \mathcal{T}_{\mathrm{conv}_0(\mathbb{1}^D)}^{<t}\right] \\ &\cdot p\left(\mathbf{p}_{n+1} = \mathbf{p} \mid t_{n+1} \in [t, t+\mathrm{d}t], \mathcal{T}_{\mathrm{conv}_0(\mathbb{1}^D)}^{<t}\right) \\ =&\bar{\lambda}\mathrm{d}t \cdot g_\tau\left(\begin{bmatrix}\mathbf{p} \\ 1 - \|\mathbf{p}\|_1\end{bmatrix}; \boldsymbol{\pi}_{\bar{\lambda}} \circ \boldsymbol{\lambda}\left(t \mid \mathcal{T}_{\mathrm{conv}_0(\mathbb{1}^D)}^{\leq t_n}\right)\right),\end{aligned}$$

where let $\mathcal{T}^{<t}_{\mathrm{conv}_0(\mathbb{1}^D)}$ denote the event $t_{n+1} \notin (t_n, t)$ and $\mathcal{T}^{\leq t_n}_{\mathrm{conv}_0(\mathbb{1}^D)}$. This suggests that in a differentiable point process, time stamps are distributed according to the homogeneous Poisson process with intensity $\bar{\lambda}$, and each mark is distributed according to the concrete distribution. □

We then investigate two properties of the differentiable point process. Proposition 5 states that a realization of the differentiable point process is differentiable with respect to model parameters under mild conditions. Proposition 6 states that the differentiable point process becomes equivalent to the original point process as temperature goes zero.

Proposition 5. *Let $\lambda_\theta(t, \mathbf{p} \mid \mathcal{T}^{\leq t_n}_{\mathbb{1}^D})$ be a conditional intensity function of a multivariate point process parameterized by θ. Assume that the conditional intensity function can be calculated with $\mathcal{T}^{\leq t_n}_{\mathrm{conv}_0(\mathbb{1}^D)}$ and is differentiable with respect to any mark in $\mathcal{T}^{\leq t_n}_{\mathrm{conv}_0(\mathbb{1}^D)}$ and θ. Then, the marks of a realization of the corresponding differentiable point process is differentiable with respect to θ.*

Proof. We prove Proposition 5 by induction. The first mark is distributed according to $\boldsymbol{\pi}_{\bar{\lambda}} \circ \boldsymbol{\lambda}_\theta\left(t \mid \emptyset\right)$, which is differentiable with respect to θ. Assume that marks observed up to (but not including) time t are differentiable with respect to θ. A mark $\mathbf{p}$ at time t is a realization of the concrete distribution with parameter $\boldsymbol{\pi} := \boldsymbol{\pi}_{\bar{\lambda}} \circ \boldsymbol{\lambda}_\theta\left(t \mid \mathcal{T}^{\leq t_n}_{\mathrm{conv}_0(\mathbb{1}^D)}\right)$, and thus, is differentiable with respect to $\boldsymbol{\pi}$. $\boldsymbol{\pi}$ is differentiable with respect to $\boldsymbol{\lambda}_\theta$, and $\boldsymbol{\lambda}_\theta$ is assumed to be differentiable with respect to θ and the past marks, which are differentiable by assumption, and this completes the proof. □

Proposition 6. *Assume that $\lambda\left(t, \mathbf{p} \mid \mathcal{T}^{\leq t_n}_{\mathrm{conv}_0(\mathbb{1}^D)}\right) = \lambda\left(t, \mathbf{p} \mid \mathcal{T}^{\leq t_n}_{\mathrm{conv}_0(\mathbb{1}^D)} \backslash \{(t_k, \mathbf{p}_k)\}\right)$ holds for any $\mathcal{T}^{\leq t_n}_{\mathrm{conv}_0(\mathbb{1}^D)}$ and any $k \in [n]$ such that $\mathbf{p}_k = \mathbf{0}$. Then, in the limit of $\tau \to +0$, the output of Algorithm 2 is distributed according to $\mathcal{MPP}(\lambda)$ if we discard the event with mark $\mathbf{p} = \mathbf{0}$.*

Proof. As proven by Maddison et al. 2017, the random variable following the concrete distribution converges to the one-hot representation of the categorical variate in the limit of $\tau \to +0$. The random variable in line 6 of Algorithm 2 satisfies the following,

$$\Pr\left[\lim_{\tau \to +0} \begin{bmatrix} \mathbf{p} \\ r \end{bmatrix} = \mathbf{1}_d \mid s, \mathcal{T}\right] = \frac{\lambda(s, \mathbf{1}_d \mid \mathcal{T})}{\bar{\lambda}}, \tag{15}$$

for $d \in [D+1]$, where let $\lambda(s, \mathbf{1}_{D+1} \mid \mathcal{T}) \equiv \bar{\lambda} - \sum_{d=1}^{D} \lambda(s, \mathbf{1}_d \mid \mathcal{T})$. The above expression states that if the value of the conditional intensity function is the same, the random variable in line 6 of Algorithm 2 is equivalent to that in line 6 of Algorithm 1. The only difference between these algorithms is whether the event with zero mark $\mathbf{p} = \mathbf{0}$ is discarded (Algorithm 1) or not (Algorithm 2). If the assumption made in Proposition 6 is satisfied, zero marks do not affect the conditional intensity function, and thus, the output of Algorithm 2 is distributed according to $\mathcal{MPP}(\lambda)$ if we discard the events with zero marks. □

D. Properties of ∂SNNs

This section describes the properties of ∂SNNs. First, Proposition 7 states that the conditional intensity function (Equation (13)) satisfies all of the conditions listed in Proposition 2, and thus, it defines an $\bar{\mathcal{N}}$-marked point process uniquely.

Proposition 7. *Assume that the filter functions $\{f_{d',d}(s)\}_{d,d' \in [D]}$ are continuous with respect to s. Then, the conditional intensity function (Equation (13)) uniquely defines an $\bar{\mathcal{N}}$-marked point process.*

Proof. We will confirm the assumptions of Proposition 2. Observing that,

$$\begin{aligned}
&\int_{t_n}^{t} \mathrm{d}s \int_{\bar{\mathcal{N}}} \mathrm{d}\mathbf{p} \lambda\left(s, \mathbf{p} \mid \mathcal{T}^{\leq t_n}_{\bar{\mathcal{N}}}\right) \\
&= \int_{t_n}^{t} \mathrm{d}s \left[\int_{\mathrm{conv}_0(\mathcal{H})} \mathrm{d}\mathbf{p} \lambda\left(s, \mathbf{p} \mid \mathcal{T}^{\leq t_n}_{\bar{\mathcal{N}}}\right) \right. \\
&\qquad \left. + \sum_{\mathbf{p} \in \mathcal{O}} \lambda\left(s, \mathbf{p} \mid \mathcal{T}^{\leq t_n}_{\bar{\mathcal{N}}}\right)\right] \\
&= \int_{t_n}^{t} \mathrm{d}s \left[\bar{\lambda} + \sum_{\mathbf{p} \in \mathcal{O}} \lambda\left(s, \mathbf{p} \mid \mathcal{T}^{\leq t_n}_{\bar{\mathcal{N}}}\right)\right] \\
&= \bar{\lambda}(t - t_n) + \int_{t_n}^{t} \mathrm{d}s \sum_{\mathbf{p} \in \mathcal{O}} \lambda\left(s, \mathbf{p} \mid \mathcal{T}^{\leq t_n}_{\bar{\mathcal{N}}}\right),
\end{aligned}$$

the first condition is satisfied. By taking $t \to \infty$ in the above expression, we can confirm that the second condition is satisfied (the first term $\bar{\lambda}(t - t_n)$ goes to infinity and the second term is guaranteed to be non-negative). The third condition is satisfied because $\lambda(s, \mathbf{p} \mid \mathcal{T}^{\leq t_n}_{\bar{\mathcal{N}}})$ is a continuous function with respect to $s > t_n$, which can be guaranteed by the continuity of the filter functions. □

Then, let us discuss two properties of ∂SNN. Proposition 8 states that the objective function is differentiable with respect to ϕ. Proposition 9 states that ∂SNN becomes equivalent to the vanilla SNN in the limit of $\tau \to +0$.

Proposition 8. *Assume that the filter functions $\{f_{d',d}(s)\}_{d,d' \in [D]}$ are differentiable with respect to their parameters ϕ. The Monte-Carlo approximation of ELBO (Equation (9)) is differentiable with respect to ϕ if we employ $\partial\mathcal{PP}(\lambda_q(t, \mathbf{p} \mid \mathcal{T}_{\bar{\mathcal{N}}}; \phi); \bar{\lambda}, \tau)$ as the variational distribution.*

Proof. We first show that marks of a realization of the variational distribution $\mathcal{T}_{\mathcal{H}}(\phi)$ are differentiable with respect to ϕ. We then show that both $\log p(\mathcal{T}_{\mathcal{O}}, \mathcal{T}_{\mathrm{conv}_0(\mathcal{H})})$ and $\log q(\mathcal{T}_{\mathrm{conv}_0(\mathcal{H})})$ are differentiable with respect to the marks of $\mathcal{T}_{\mathcal{H}}(\phi)$. We finally show that $\log q(\mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}; \phi)$ is differentiable with respect to ϕ. By the fact that the composition of differentiable functions is differentiable, the proposition is implied by these three statements.

$\mathcal{T}_{\mathcal{H}}(\phi)$ is sampled by Algorithm 2 using the conditional intensity function $\lambda_q(t, \mathbf{p} \mid \mathcal{T}_{\bar{\mathcal{N}}}; \phi)$, which is differentiable with respect to any mark in $\mathcal{T}_{\bar{\mathcal{N}}}$ and ϕ (by the assumption). Therefore, by Proposition 5, the marks of $\mathcal{T}_{\mathcal{H}}(\phi)$ are differentiable with respect to ϕ.

The logarithm of the joint distribution $\log p(\mathcal{T}_{\mathcal{O}}, \mathcal{T}_{\mathrm{conv}_0(\mathcal{H})})$ can be written as,

$$\begin{aligned}
&\log p(\mathcal{T}_{\mathcal{O}}, \mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}; \theta) \\
&= \sum_{(t,\mathbf{p}) \in \mathcal{T}_{\mathcal{O}}} \log \lambda^{\mathrm{SNN}} \left(t, \mathbf{p} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_{n(t)}} \right) \\
&\quad + \sum_{(t,\mathbf{p}) \in \mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}} \log \lambda_{\partial} \left(t, \mathbf{p}_{\mathcal{H}} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_{n(t)}}; \boldsymbol{\lambda}_{\mathcal{H}}, \bar{\lambda}, \tau \right) \\
&\quad - \bar{\lambda} T - \int_0^T \mathrm{d}s \sum_{\mathbf{p} \in \mathcal{O}} \lambda^{\mathrm{SNN}} \left(s, \mathbf{p} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_{n(s)}} \right).
\end{aligned}$$

The first and the last terms are differentiable with respect to marks of $\mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}$, and the third term does not depend on $\mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}$. For any $(t, \mathbf{p}) \in \mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}$, the summand of the second term,

$$\begin{aligned}
&\log \lambda_{\partial} \left(t, \mathbf{p}_{\mathcal{H}} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_{n(t)}}; \boldsymbol{\lambda}_{\mathcal{H}}, \bar{\lambda}, \tau \right) \\
&= \log \bar{\lambda} + \log g_{\tau} \left(\begin{bmatrix} \mathbf{p}_{\mathcal{H}} \\ 1 - \|\mathbf{p}_{\mathcal{H}}\|_1 \end{bmatrix}; \boldsymbol{\pi}_{\bar{\lambda}} \circ \boldsymbol{\lambda}_{\mathcal{H}} \left(t \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_n(t)} \right) \right),
\end{aligned}$$

is differentiable with respect to $\mathbf{p}_{\mathcal{H}}$ because the probability density function of the concrete distribution is differentiable with respect to $\mathbf{p}_{\mathcal{H}}$. It is also differentiable with respect to the past marks in $\mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_{n(t)}}$ because the probability density function g_{τ} is differentiable with respect to its parameter $\boldsymbol{\pi} := \boldsymbol{\pi}_{\bar{\lambda}} \circ \boldsymbol{\lambda}_{\mathcal{H}} \left(t \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_n(t)} \right)$, which is differentiable with respect to the marks in $\mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_n(t)}$. Therefore, $\log p(\mathcal{T}_{\mathcal{O}}, \mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}; \theta)$, is differentiable with respect to the marks of $\mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}$.

The logarithm of the variational distribution $\log q(\mathcal{T}_{\mathrm{conv}_0(\mathcal{H})})$ can be written as,

$$\begin{aligned}
&\log q(\mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}) \\
&= \sum_{(t,\mathbf{p}) \in \mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}} \log \lambda_{\partial} \left(t, \mathbf{p}_{\mathcal{H}} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_{n(t)}}; \boldsymbol{\lambda}_{\mathcal{H}}, \bar{\lambda}, \tau \right) - \bar{\lambda} T,
\end{aligned} \tag{16}$$

which is also differentiable with respect to the marks of $\mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}$ in the same way as the above discussion.

Finally, $\log q(\mathcal{T}_{\mathrm{conv}_0(\mathcal{H})}; \phi)$ is differentiable with respect to ϕ, because the first term of Equation (16) is differentiable with respect to $\boldsymbol{\lambda}_{\mathcal{H}}$, which is differentiable with respect to ϕ (partially by the assumption that the filter functions are differentiable with respect to ϕ).

The proposition follows from the three statements above. □

Proposition 9. *In the limit of $\tau \to +0$, a realization of ∂SNN (Equation (13)) is distributed according to the vanilla SNN (Equation (7)) if we discard events with mark $\mathbf{p} = \mathbf{0}$.*

Proof. Since for any event $(t_k, \mathbf{p}_k)$ $(k \in [n])$ such that $\mathbf{p}_k = \mathbf{0}$,

$$\lambda^{\mathrm{SNN}} \left(t, \mathbf{p} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_n} \right) = \lambda^{\mathrm{SNN}} \left(t, \mathbf{p} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_n} \backslash \{(t_k, \mathbf{p}_k)\} \right),$$

holds, the event with mark $\mathbf{p} = \mathbf{0}$ has no influence on the conditional intensity function. In the limit of $\tau \to +0$,

$$\begin{aligned}
&\lambda^{\partial \mathrm{SNN}} \left(t, \mathbf{p} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_n} \right) \\
&= \begin{cases} \lambda^{\mathrm{SNN}} \left(t, \mathbf{p} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_n} \right) & (\mathbf{p} \in \mathcal{O} \cup \mathcal{H}), \\ \left(\bar{\lambda} - \sum_{\mathbf{p} \in \mathcal{H}} \lambda^{\mathrm{SNN}} \left(t, \mathbf{p} \mid \mathcal{T}_{\bar{\mathcal{N}}}^{\leq t_n} \right) \right) & (\mathbf{p} = \mathbf{0}), \end{cases}
\end{aligned}$$

holds. As discussed above, the event with mark $\mathbf{p} = \mathbf{0}$ has no influence on computing the conditional intensity function, and can be removed without changing the conditional intensity function. Therefore, a realization of $\mathcal{MPP}(\lambda^{\partial \mathrm{SNN}})$ is equivalent to that of $\mathcal{MPP}(\lambda^{\mathrm{SNN}})$ if we discard all the events with mark $\mathbf{p} = \mathbf{0}$. □

E. Numerically Stable Implementation

For numerical stability, we recommend to represent all probability and subprobability vectors and conditional intensity functions in the logarithmic scale. In this section, we describe an accurate computation of the parameter of the concrete distribution.

When computing the parameter of the concrete distribution shown below in the logarithmic scale,

$$\begin{aligned}
&\boldsymbol{\pi}_{\bar{\lambda}} \circ \boldsymbol{\lambda} \left(t \mid \mathcal{T}_{\mathrm{conv}_0(\mathbb{1}^D)}^{\leq t_n} \right) \\
&= \frac{1}{\bar{\lambda}} \left[\boldsymbol{\lambda} \left(t \mid \mathcal{T}_{\mathrm{conv}_0(\mathbb{1}^D)}^{\leq t_n} \right) \quad \bar{\lambda} - \left\| \boldsymbol{\lambda} \left(t \mid \mathcal{T}_{\mathrm{conv}_0(\mathbb{1}^D)}^{\leq t_n} \right) \right\|_1 \right],
\end{aligned}$$

it is straightforward to compute the first D elements with the logarithm of the conditional intensity function,

$\log \lambda \left(t, \mathbf{p} \mid \mathcal{T}^{<t}_{\mathrm{conv}_0(\mathbb{1}^D)}\right)$. However, the last element,

$$1 - \frac{\sum_{\mathbf{p}' \in \mathbb{1}^D} \lambda \left(t, \mathbf{p}' \mid \mathcal{T}^{<t}_{\mathrm{conv}_0(\mathbb{1}^D)}\right)}{\bar{\lambda}}, \tag{17}$$

is not trivial to compute accurately in the logarithmic scale.

We resort to `log1mexp` (Mächler, 2012) to compute it, which allows us to compute $\log(1 - \exp(-|a|))$ accurately for $a \neq 0$ as follows:

$$\texttt{log1mexp}(a) = \begin{cases} \log(-\texttt{expm1}(-a)) & (0 < a \leq \log 2), \\ \texttt{log1p}(-\exp(-a)) & (a > \log 2), \end{cases}$$

where $\texttt{expm1}(\texttt{x})$ and $\texttt{log1p}(x)$ approximately compute $\exp(x) - 1$ and $\log(1 + x)$ respectively by using a few terms of their Taylor series. Since we can compute $\log p := \log \left[\sum_{\mathbf{p}' \in \mathbb{1}^D} \lambda \left(t, \mathbf{p}' \mid \mathcal{T}^{<t}_{\mathrm{conv}_0(\mathbb{1}^D)}\right)\right] - \log \bar{\lambda}$ by `logsumexp`, the computation of Equation (17) boils down to the computation of $\log(1-p)$ given $\log p$ ($p \in [0, 1)$). Since $\log(1 - p) = \log(1 - \exp(-|\log p|))$ holds for $p \in [0, 1)$, we can utilize `log1mexp` to compute it.