Thinning for Accelerating the Learning of Point Processes (Supplementary Material)

Lemma 3.2 (Thinned intensities)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be the full history and thinned history with respect to a p-thinned process $N_p(t)$. Let $\mathcal{H}$ be the internal history of $N(t)$. The following equalities hold:*

(1) $\lambda_p^{\mathcal{F}}(t) = p\lambda^{\mathcal{H}}(t)$*;*

(2) $\lambda_p^{\mathcal{G}}(t) = p\mathbb{E}\left[\lambda^{\mathcal{H}}(t)|\mathcal{G}\right]$.

Proof. For (1), it can be obtained by taking expectation on both side of $dN_p(t) = B_{N(t)}dN(t)$:

$$\lambda_p^{\mathcal{F}}(t) = \mathbb{E}dN_p(t) = \mathbb{E}B_{N(t)}dN(t) = \lambda^{\mathcal{H}}(t). \tag{1}$$

For (2), Theorem 7.13 in [2] gives a solution to recover the point process given the thinned history be the following conditional expectation:

$$\mathbb{E}\left[N(t)|\mathcal{G}\right] = N_p(t) + \frac{1-p}{p}\int_0^t d\Lambda_p^{\mathcal{G}}(s). \tag{2}$$

Here, $\Lambda_p^{\mathcal{G}}$ is the $\mathcal{G}$-compensator of the p-thinned process, which equals to $\Lambda_p^{\mathcal{G}}(t) = \int_0^t \lambda_p^{\mathcal{G}}(s)ds$. Further,

$$\begin{aligned}
\mathbb{E}\left[\lambda^{\mathcal{H}}(t)|\mathcal{G}\right] &= \lim_{s\to 0}\frac{\mathbb{E}\left[N(t+s)-N(t)|\mathcal{G}\right]}{ds} \\
&= \lim_{s\to 0}\frac{\mathbb{E}\left[N_p(t+s)-N_p(t)|\mathcal{G}\right]}{ds} + \frac{1-p}{p}\lambda_p^{\mathcal{G}}(t) \\
&= \lambda_p^{\mathcal{G}}(t) + \frac{1-p}{p}\lambda_p^{\mathcal{G}}(t) \\
&= \frac{1}{p}\lambda_p^{\mathcal{G}}(t).
\end{aligned}$$

where the desired result follows. □

Lemma 4.1 (Thinning for parameter estimation of NHPP)**.** *Consider an NHPP $N(t)$ with deterministic intensity $\lambda(t;\theta)$, $t > 0$, $\theta \in \mathbb{R}^d$. If there exists an invertible linear operator $\mathcal{A}: \mathbb{R}^d \to \mathbb{R}^d$ satisfying $\lambda(t;\mathcal{A}\theta) = p\lambda(t;\theta)$, then the M-estimator on thinned history can be written as $\tilde{\theta}_{\mathcal{H}} = \mathcal{A}^{-1}\hat{\theta}_{\mathcal{G}}$ such that $\mathbb{E}\left[\nabla R(\tilde{\theta}_{\mathcal{H}})|\mathcal{G}\right] \xrightarrow{P} 0$, as the number of realizations $n \to \infty$.*

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Proof. From Theorem 7.13 in [2] we have,

$$\begin{aligned}\mathbb{E}\big[\nabla R(\tilde\theta_{\mathcal{H}})|\mathcal{G}\big] &= \mathbb{E}\Big\{\frac{1}{p}\int_0^T H(t;\tilde\theta_{\mathcal{H}})\,\Big[dN_p(t) - p\lambda(t;\tilde\theta_{\mathcal{H}})dt\Big|\mathcal{G}\Big]\,\Big\}\\ &= \frac{1}{p}\int_0^T H(t;\tilde\theta_{\mathcal{H}})\,\big[dN_p(t) - \lambda(t;\hat\theta_{\mathcal{G}})dt\big]\\ &= \frac{1}{p}\int_0^T H(t;\tilde\theta_{\mathcal{H}})\,\big[dN_p(t) - \lambda(t;\theta^*_{\mathcal{G}})dt + \lambda(t;\theta^*_{\mathcal{G}})dt - \lambda(t;\hat\theta_{\mathcal{G}})dt\big]\\ &= \frac{1}{p}\int_0^T H(t;\tilde\theta_{\mathcal{H}})\,\big[\lambda(t;\theta^*_{\mathcal{G}})dt - \lambda(t;\hat\theta_{\mathcal{G}})dt\big] \xrightarrow{P} 0.\end{aligned}$$

The last step is due to the asymptotic normality of M-estimator ([1]) that $\hat\theta_{\mathcal{G}} \xrightarrow{P} \theta^*_{\mathcal{G}}$ as the number of realizations $n \to \infty$. □

Theorem 4.3 (Thinning for parameter estimation of decouplable intensities)**.** *Consider a point process* $N(t)$ *with decouplable intensity. If there exist invertible linear operators* $\mathcal{A}$ *and* $\mathcal{B}$ *satisfying* $\mathcal{B}\mathbb{E}\big[m^{\mathcal{H}}(t)|\mathcal{G}\big] = m_p^{\mathcal{G}}(t)$*, where* $m_p^{\mathcal{G}}(t)$ *is the component of thinned intensity* $\lambda_p^{\mathcal{G}}(t)$*, and* $p\mathcal{B}^{-1}g(t;\theta) = g(t;\mathcal{A}\theta)$*, then the M-estimator on thinned history can be written as* $\tilde\theta_{\mathcal{H}} = \mathcal{A}^{-1}\hat\theta_{\mathcal{G}}$ *such that* $\mathbb{E}\big[\nabla R(\tilde\theta_{\mathcal{H}})|\mathcal{G}\big] \xrightarrow{P} 0$*, as the number of realizations* $n \to \infty$*. Particularly, if* $\lambda^{\mathcal{H}}(t;\theta)$ *is linear, then* $\mathcal{A} = p\mathcal{B}^{-1}$.

Proof. The proof is similar with the NHPP one. Be definition we have,

$$\begin{aligned}\mathbb{E}\big[\nabla R(\tilde\theta_{\mathcal{H}})|\mathcal{G}\big] &= \frac{1}{p}\mathbb{E}\int_0^T \Big\{H^{\mathcal{H}}(t;\tilde\theta_{\mathcal{H}})\big[dN_p(t) - p\lambda^{\mathcal{H}}(t;\tilde\theta_{\mathcal{H}})dt\big]\Big|\mathcal{G}\Big\}\\ &= \frac{1}{p}\int_0^T \mathbb{E}\left\{H^{\mathcal{H}}(t;\tilde\theta_{\mathcal{H}})|\mathcal{G}\right\}\mathbb{E}\Big\{dN_p(t) - p\lambda^{\mathcal{H}}(t;\tilde\theta_{\mathcal{H}})dt\Big|\mathcal{G}\Big\}\end{aligned}$$

By the definition of stochastic integral, it suffices to show that $N_p(t) - \int p\lambda^{\mathcal{H}}(t;\tilde\theta_{\mathcal{H}})dt$ asymptotically converges to a martingale in probability.

$$\begin{aligned}\mathbb{E}\Big\{dN_p(t) - p\lambda^{\mathcal{H}}(t;\tilde\theta_{\mathcal{H}})dt\Big|\mathcal{G}\Big\} &= g(t;\theta^*)^T m^{\mathcal{G}}(t) - pg(t;\tilde\theta_{\mathcal{H}})^T\mathbb{E}\left[m^{\mathcal{H}}(t)|\mathcal{G}\right]\\ &= \left[g(t;\theta^*) - p\mathcal{B}^{-1}g(t;\tilde\theta_{\mathcal{H}})\right]^T m^{\mathcal{G}}(t)\\ &= \left[g(t;\theta^*) - g(t;\mathcal{A}\tilde\theta_{\mathcal{H}})\right]^T m^{\mathcal{G}}(t)\\ &= \left[g(t;\theta^*) - g(t;\hat\theta_{\mathcal{G}})\right]^T m^{\mathcal{G}}(t)\end{aligned}$$

Since $\hat\theta_{\mathcal{G}} \xrightarrow{P} \theta^*_{\mathcal{G}}$ as the number of realizations $n \to \infty$, and g is continuous with respect to θ, $\left[g(t;\theta^*) - g(t;\hat\theta_{\mathcal{G}})\right]^T m^{\mathcal{G}}(t) \xrightarrow{P} 0$, which is the desired result. □

Theorem 5.1 (Thinning for gradient estimation)**.** *Let* $N(t)$ *be a point process with decouplable intensity* $\lambda^{\mathcal{H}}(t;\theta) = g(t;\theta)^T m^{\mathcal{H}}(t)$ *in Eq. (4). If there exist invertible linear operators* $\mathcal{A}$ *and* $\mathcal{B}$ *satisfying* $\mathcal{B}\mathbb{E}\big[m^{\mathcal{H}}(t)|\mathcal{G}\big] = m_p^{\mathcal{G}}(t)$*, where* $m_p^{\mathcal{G}}(t)$ *is the component of thinned intensity* $\lambda_p^{\mathcal{G}}(t)$*, and* $p\mathcal{B}^{-1}g(t;\theta) = g(t;\mathcal{A}\theta)$*, then*

(1) $\mathbb{E}\big[\nabla R(\theta)|\mathcal{G}\big] \leqslant 1/p\mathcal{A}^{-1}\nabla R_p(\mathcal{A}\theta)$*, for* R *is LSE;*

(2) $\mathbb{E}\big[\nabla R(\theta)|\mathcal{G}\big] \leqslant \mathcal{A}^{-1}\nabla R_p(\mathcal{A}\theta)$*, for* R *is MLE.*

Particularly, if the intensity is deterministic, i.e., $m^{\mathcal{H}}(t) = 1$*, both equalities hold.*

Proof. By definition of stochastic integral, we have

$$\begin{aligned}\mathbb{E}\big[\nabla R(\theta)|\mathcal{G}\big] &= \mathbb{E}\left\{\int_0^T H^{\mathcal{H}}(t;\theta)\left[dN(t) - \lambda^{\mathcal{H}}(t;\theta)dt\right]\Big|\mathcal{G}\right\}\\ &= \mathbb{E}\left\{\int_0^T H^{\mathcal{H}}(t;\theta)dN(t)\Big|\mathcal{G}\right\} - \mathbb{E}\left\{\int_0^T H^{\mathcal{H}}(t;\theta)\lambda^{\mathcal{H}}(t;\theta)dt\Big|\mathcal{G}\right\}\end{aligned}$$

Here the second term can be bounded by,

$$\mathbb{E}\left\{H^{\mathcal{H}}(t;\theta)\lambda^{\mathcal{H}}(t;\theta)dt\middle|\mathcal{G}\right\} \geqslant \mathbb{E}\left\{H^{\mathcal{H}}(t;\theta)\middle|\mathcal{G}\right\}\mathbb{E}\left\{\lambda^{\mathcal{H}}(t;\theta)dt\middle|\mathcal{G}\right\}$$

According to the definition of forward stochastic integral, the first term can be written as,

$$\mathbb{E}\left\{\int_0^T H^{\mathcal{H}}(t;\theta)dN(t)|\mathcal{G}\right\} = \int_0^T \mathbb{E}\left\{H^{\mathcal{H}}(t;\theta)|\mathcal{G}\right\}\mathbb{E}\{dN(t)|\mathcal{G}\}$$

Let's look at these components one by one. The condition of the theorem yields,

$$\begin{aligned}\mathbb{E}\left\{\lambda^{\mathcal{H}}(t;\theta)dt\middle|\mathcal{G}\right\} &= g(t;\theta)^{\top}\mathbb{E}\big[m^{\mathcal{H}}(t)|\mathcal{G}\big]dt\\ &= p\mathcal{B}^{-1}g(t;\theta)^{\top}m^{\mathcal{G}}(t)dt\\ &= g(t;\mathcal{A}\theta)^{\top}m^{\mathcal{G}}(t)dt\\ &= \lambda^{\mathcal{G}}(t;\mathcal{A}\theta)dt \qquad (3)\end{aligned}$$

and,

$$\mathbb{E}\{dN(t)|\mathcal{G}\} = \frac{1}{p}dN_p(t). \qquad (4)$$

If R is LSE, then we have,

$$\begin{aligned}\mathbb{E}\Big[H^{\mathcal{H}}(t;\theta)\Big|\mathcal{G}\Big] &= \nabla\mathbb{E}\Big[\lambda^{\mathcal{H}}(t;\theta)|\mathcal{G}\Big]\\ &= \nabla_{\theta}\frac{1}{p}g(t;\mathcal{A}\theta)^{\top}m_p^{\mathcal{G}}(t)\\ &= \mathcal{A}^{-1}\nabla\frac{1}{p}g(t;\mathcal{A}\theta)^{\top}m_p^{\mathcal{G}}(t)\\ &= \mathcal{A}^{-1}H_p^{\mathcal{G}}(t;\mathcal{A}\theta) \qquad (5)\end{aligned}$$

Thus, combining Eq.(3),(4) and (5) yields,

$$\begin{aligned}\mathbb{E}\big[\nabla R(\theta)|\mathcal{G}\big] &\leqslant \frac{1}{p^2}\mathcal{A}^{-1}\int_0^T H_p^{\mathcal{G}}(t;\mathcal{A}\theta)\big[dN_p(t)-\lambda_p^{\mathcal{G}}(t;\mathcal{A}\theta)dt\big]\\ &= \frac{1}{p}\mathcal{A}^{-1}\nabla R_p(\mathcal{A}\theta).\end{aligned}$$

If R is MSE, then we have,

$$\begin{aligned}\mathbb{E}\Big[H^{\mathcal{H}}(t;\theta)\Big|\mathcal{G}\Big] &= \nabla\mathbb{E}\Big[\log\lambda^{\mathcal{H}}(t;\theta)|\mathcal{G}\Big]\\ &\geqslant \nabla_{\theta}\log\Big[g(t;\mathcal{A}\theta)^{\top}m_p^{\mathcal{G}}(t)\Big]\\ &= \mathcal{A}^{-1}\nabla\log\Big[g(t;\mathcal{A}\theta)^{\top}m_p^{\mathcal{G}}(t)\Big]\\ &= \mathcal{A}^{-1}H_p^{\mathcal{G}}(t;\mathcal{A}\theta) \qquad (6)\end{aligned}$$

Combining Eq.(3) and Eq.(6) yields the second conclusion,

$$\begin{aligned}\mathbb{E}\big[\nabla R(\theta)|\mathcal{G}\big] &\leqslant \frac{1}{p}\mathcal{A}^{-1}\int_0^T H_p^{\mathcal{G}}(t;\mathcal{A}\theta)\big[dN_p(t)-\lambda_p^{\mathcal{G}}(t;\mathcal{A}\theta)dt\big]\\ &= \mathcal{A}^{-1}\nabla R_p(\mathcal{A}\theta).\end{aligned}$$

The proof ends here. □

Theorem 5.2 (Variance of gradient estimation)**.** *Let* $\nabla\hat{R}^{\mathcal{G}}(\theta)$ *and* $\nabla R_{\ell}(\theta)$ *be the* p*-thinned and sub-interval gradient at* θ*, where* $\nabla\hat{R}^{\mathcal{G}}(\theta) = 1/p\mathcal{A}^{-1}\nabla R_p(\mathcal{A}\theta)$ *for LSE and* $\nabla\hat{R}^{\mathcal{G}}(\theta) = \mathcal{A}^{-1}\nabla R_p(\mathcal{A}\theta)$ *for MLE. The variance of* p*-thinned gradient is no greater than that of sub-interval gradient:*

$$\mathbb{V}\Big[\nabla\hat{R}^{\mathcal{G}}(\theta)\Big] \leqslant \mathbb{V}\Big[\nabla R_{\ell}(\theta)\Big]. \qquad (7)$$

Proof. For the RHS, using the law of total variance yields,

$$\mathbb{V}\Big[\nabla R_\ell(\theta)\Big] = \mathbb{E}\Big\{\mathbb{V}\big[\nabla R_\ell(\theta)|\mathcal{F}\big]\Big\} + \mathbb{V}\Big\{\mathbb{E}\big[\nabla R_\ell(\theta)|\mathcal{F}\big]\Big\}.$$

The first term can be rewritten as,

$$\begin{aligned}\mathbb{E}\Big\{\mathbb{V}\big[\nabla R_\ell(\theta)|\mathcal{F}\big]\Big\} &= \frac{1-p}{p}\mathbb{E}\left\{\int_0^T H^{\mathcal{H}}(t;\theta)\big[dN(t) - \lambda^{\mathcal{H}}(t;\theta)dt\big]\right\}^2 \\ &= \frac{1-p}{p}\mathbb{E}\Big[\nabla R(\theta)\Big]^2,\end{aligned}$$

The second term can be written as,

$$\mathbb{V}\Big\{\mathbb{E}\big[\nabla R_\ell(\theta)|\mathcal{F}\big]\Big\} = \mathbb{V}\big[\nabla R(\theta)\big].$$

Thus, the total variance of $\nabla R_\ell(\theta)$ can be written as,

$$\mathbb{V}\Big[\nabla R_\ell(\theta)\Big] = \frac{1-p}{p}\mathbb{E}\Big[\nabla R(\theta)\Big]^2 + \mathbb{V}\big[\nabla R(\theta)\big].$$

Then we consider the LHS, by the definition of variance,

$$\mathbb{V}\Big[\nabla \hat{R}^{\mathcal{G}}(\theta)\Big] = \mathbb{E}\Big[\nabla \hat{R}^{\mathcal{G}}(\theta)\Big]^2 - \Big[\mathbb{E}\nabla \hat{R}^{\mathcal{G}}(\theta)\Big]^2. \tag{8}$$

Apply Theorem **??**, we have,

$$\Big[\mathbb{E}\nabla \hat{R}^{\mathcal{G}}(\theta)\Big]^2 \geqslant \Big[\mathbb{E}\nabla R(\theta)\Big]^2. \tag{9}$$

For LSE, since quadratic function is convex, we obtain $\mathbb{E}\big[H_p^{\mathcal{G}}(t;\mathcal{A}\theta)\big]^2 \leqslant \mathbb{E}\big[H^{\mathcal{H}}(t;\theta)\big]^2$. This equivalence also holds for MLE, we omit the proof, since it can be proved similarly. Further, we obtain,

$$\begin{aligned}\mathbb{E}\Big[\nabla \hat{R}^{\mathcal{G}}(\theta)\Big]^2 =& \frac{1}{p^2}\mathcal{A}^{-1}\mathbb{E}\Big[\int_0^T H_p^{\mathcal{G}}(t;(\mathcal{A}\theta)\left[dN_p(t) - \lambda_p^{\mathcal{G}}(t;\mathcal{A}\theta)dt\right]\Big]^2(\mathcal{A}^{-1})^\top \\ =& \frac{1}{p^2}\mathcal{A}^{-1}\mathbb{E}\Big[\int_0^T H_p^{\mathcal{G}}(t;(\mathcal{A}\theta)\left[dN_p(t) - \lambda_p^{\mathcal{G}}(t;\theta^*_{\mathcal{G}})dt + \lambda_p^{\mathcal{G}}(t;\theta^*_{\mathcal{G}})dt - \lambda_p^{\mathcal{G}}(t;\mathcal{A}\theta)dt\right]\Big]^2(\mathcal{A}^{-1})^\top \\ =& \frac{1}{p^2}\mathcal{A}^{-1}\mathbb{E}\left\{\int_0^T H_p^{\mathcal{G}}(t;\mathcal{A}\theta)\left[dN_p(t) - \lambda_p^{\mathcal{G}}(t;\theta^*_{\mathcal{G}})dt\right]\right\}^2(\mathcal{A}^{-1})^\top + \\ & \frac{1}{p^2}\mathcal{A}^{-1}\mathbb{E}\left\{\int_0^T H_p^{\mathcal{G}}(t;\mathcal{A}\theta)\left[\lambda_p^{\mathcal{G}}(t;\mathcal{A}\theta^*_{\mathcal{G}}) - \lambda_p^{\mathcal{G}}(t;\mathcal{A}\theta)\right]dt\right\}^2(\mathcal{A}^{-1})^\top\end{aligned} \tag{10}$$

The first term,

$$\begin{aligned}& \frac{1}{p^2}\mathcal{A}^{-1}\mathbb{E}\left\{\int_0^T H_p^{\mathcal{G}}(t;\mathcal{A}\theta)\left[dN_p(t) - \lambda_p^{\mathcal{G}}(t;\theta^*_{\mathcal{G}})dt\right]\right\}^2(\mathcal{A}^{-1})^\top \\ =& \frac{1}{p^2}\mathcal{A}^{-1}\mathbb{E}\left\{\int_0^T \left[H_p^{\mathcal{G}}(t;\mathcal{A}\theta)\right]^2 dN_p(t)\right\}(\mathcal{A}^{-1})^\top \\ \leqslant& \frac{1}{p^2}\mathbb{E}\left\{\int_0^T \left[H^{\mathcal{H}}(t;\theta)\right]^2 dN_p(t)\right\} \\ =& \frac{1}{p}\mathbb{E}\int_0^T \left[H^{\mathcal{H}}(t;\theta)\right]^2 dN(t)\end{aligned} \tag{11}$$

The second term,

$$\begin{aligned}& \frac{1}{p^2}\mathcal{A}^{-1}\mathbb{E}\left\{\int_0^T H_p^{\mathcal{G}}(t;\mathcal{A}\theta)\left[\lambda_p^{\mathcal{G}}(t;\mathcal{A}\theta^*_{\mathcal{G}}) - \lambda_p^{\mathcal{G}}(t;\mathcal{A}\theta)\right]dt\right\}^2(\mathcal{A}^{-1})^\top \\ \leqslant& \frac{1}{p}\mathbb{E}\left\{\int_0^T H^{\mathcal{H}}(t;\theta)\left[\lambda^{\mathcal{H}}(t;\theta^*_{\mathcal{H}}) - \lambda^{\mathcal{H}}(t;\theta)\right]dt\right\}^2\end{aligned} \tag{12}$$

Substituting Eq. (11) and (12)to Eq. 10 yields

$$\begin{aligned}
\mathbb{E}\Big[\nabla\hat{R}^{\mathcal{G}}(\theta)\Big]^2 &\leqslant \frac{1}{p}\mathbb{E}\int_0^T \left[H^{\mathcal{H}}(t;\theta)\right]^2 dN(t) + \frac{1}{p}\mathbb{E}\left\{\int_0^T H^{\mathcal{H}}(t;\theta)\left[\lambda^{\mathcal{H}}(t;\theta^*_{\mathcal{H}}) - \lambda^{\mathcal{H}}(t;\theta)\right] dt\right\}^2 \\
&= \frac{1}{p}\mathbb{E}\left\{\int_0^T H^{\mathcal{H}}(t;\theta)\left[dN(t) - \lambda^{\mathcal{H}}(t;\theta^*_{\mathcal{H}})dt\right]\right\}^2 \\
&= \mathbb{E}\Big[\nabla R(\theta)\Big]^2. \qquad (13)
\end{aligned}$$

Combine Eq.(13) and Eq.(10) to Eq.(8),

$$\begin{aligned}
\mathbb{V}\Big[\nabla\hat{R}^{\mathcal{G}}(\theta)\Big] &\leqslant \frac{1}{p}\mathbb{E}\Big[\nabla R(\theta)\Big]^2 - \Big[\mathbb{E}\nabla R(\theta)\Big]^2 \\
&= \frac{1-p}{p}\mathbb{E}\Big[\nabla R(\theta)\Big]^2 + \mathbb{E}\Big[\nabla R(\theta)\Big]^2 - \Big[\mathbb{E}\nabla R(\theta)\Big]^2 \\
&= \frac{1-p}{p}\mathbb{E}\Big[\nabla R(\theta)\Big]^2 + \mathbb{V}\big[\nabla R(\theta)\big] \\
&= \mathbb{V}\Big[\nabla R_\ell(\theta)\Big],
\end{aligned}$$

which is the desired result. □

References

[1] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[2] Alan Karr. *Point Processes and Their Statistical Inference*, volume 7. CRC Press, 1991.