Stochastic Gradient Kernel Density Mode-Seeking (Supplemental Document)

Xiao-Tong Yuan
NLPR, CASIA
xtyuan@nlpr.ia.ac.cn

Stan Z. Li
CBSR&NLPR, CASIA
szli@nlpr.ia.ac.cn

1. Proof of Proposition 4

The technical proof below is based on the skills presented in [1].

Proof First we introduce

$$\tilde{\mathcal{R}}_N = \sum_{n=1}^{N} k_n(\hat{\mathbf{y}}^o_{n-1}) - \max_{\mathbf{y}} \left(\sum_{n=1}^{N} k_n(\mathbf{y}) \right)$$

which obviously satisfies $\mathcal{R}_N \propto \tilde{\mathcal{R}}_N$.

By concavity assumption on $k(\cdot)$, for any $\mathbf{y}^* \in \mathbb{R}^D$,

$$\begin{aligned}
& k_n(\mathbf{y}^o_{n-1}) - k_n(\mathbf{y}^*) \\
= \;& k(M^2(\mathbf{y}^o_{n-1}, \mathbf{x}_n, \mathbf{\Sigma})) - k(M^2(\mathbf{y}^*, \mathbf{x}_n, \mathbf{\Sigma})) \\
\geqslant \;& k^{'}(M^2(\mathbf{y}^o_{n-1}, \mathbf{x}_n, \mathbf{\Sigma}))(M^2(\mathbf{y}^o_{n-1}, \mathbf{x}_n, \mathbf{\Sigma}) - M^2(\mathbf{y}^*, \mathbf{x}_n, \mathbf{\Sigma})) \\
= \;& -2k^{'}(M^2(\mathbf{y}^o_{n-1}, \mathbf{x}_n, \mathbf{\Sigma}))(\mathbf{y}^o_{n-1} - \mathbf{x}_n)^T \mathbf{\Sigma}^{-1}(\mathbf{y}^* - \mathbf{y}^o_{n-1}) \\
& -k^{'}(M^2(\mathbf{y}^o_{n-1}, \mathbf{x}_n, \mathbf{\Sigma}))M^2(\mathbf{y}^o_{n-1}, \mathbf{y}^*, \mathbf{\Sigma}) \\
= \;& \frac{(\mathbf{y}^o_n - \mathbf{y}^o_{n-1})^T \mathbf{\Sigma}^{-1}(\mathbf{y}^o_{n-1} - \mathbf{y}^*)}{\eta_n} + g_n M^2(\mathbf{y}^o_{n-1}, \mathbf{y}^*, \mathbf{\Sigma})
\end{aligned}$$

where $g_n = -k^{'}(M^2(\mathbf{y}^o_{n-1}, \mathbf{x}_n, \mathbf{\Sigma}))$. Let $\mathbf{y}^* = \arg\max_{\mathbf{y}} \left(\sum_{n=1}^{N} k_n(\mathbf{y}) \right)$. Summing over n on both sides of above inequality, and recalling that $\eta_n = \left(2 \sum_{s=1}^{n} g_s\right)^{-1}$, we get

$$\begin{aligned}
& \tilde{R}_N \geqslant \sum_{n=1}^{N} \frac{(\mathbf{y}^o_n - \mathbf{y}^o_{n-1})\mathbf{\Sigma}^{-1}(\mathbf{y}^o_{n-1} - \mathbf{y}^*)}{\eta_n} \\
& + \sum_{n=1}^{N} g_n M^2(\mathbf{y}^o_{n-1}, \mathbf{y}^*, \mathbf{\Sigma}) = \\
& \sum_{n=1}^{N} \frac{1}{2\eta_n}(-M^2(\mathbf{y}^o_{n-1}, \mathbf{y}^*, \mathbf{\Sigma}) + M^2(\mathbf{y}^o_n, \mathbf{y}^*, \mathbf{\Sigma}) \\
& -M^2(\mathbf{y}^o_n, \mathbf{y}^o_{n-1}, \mathbf{\Sigma})) + \sum_{n=1}^{N} g_n M^2(\mathbf{y}^o_{n-1}, \mathbf{y}^*, \mathbf{\Sigma}) \\
& = \frac{1}{2} \sum_{n=2}^{N} M^2(\mathbf{y}^o_{n-1}, \mathbf{y}^*, \mathbf{\Sigma}) \left(\frac{1}{\eta_{n-1}} - \frac{1}{\eta_n} + 2g_n \right) \\
& + \left(-\frac{1}{\eta_1} + 2w_1 g_1 \right) M^2(\mathbf{y}^o_0, \mathbf{y}^*, \mathbf{\Sigma}) - \frac{1}{2} \sum_{n=1}^{N} \frac{1}{\eta_n} M^2(\mathbf{y}^o_n, \mathbf{y}^o_{n-1}, \mathbf{\Sigma}) \\
& = -\frac{1}{2} \sum_{n=1}^{N} \frac{1}{\eta_n} M^2(\mathbf{y}^o_n, \mathbf{y}^o_{n-1}, \mathbf{\Sigma}) \\
& = -\frac{1}{2} \sum_{n=1}^{N} \eta_n M^2(k^{'}_n(\mathbf{y}^o_{n-1}), \mathbf{0}, \mathbf{\Sigma}^{-1}) \geqslant -\frac{1}{2} \sum_{n=1}^{N} \eta_n G^2.
\end{aligned}$$

Thus we have

$$\mathcal{R}_N \geq -\frac{1}{2NC_k} \sum_{n=1}^{N} \eta_n G^2.$$

Moreover, if $g(\cdot) \geq L_1 > 0$, then $\eta_n \leq \frac{1}{2nL_1}$ and

$$\begin{aligned}
\sum_{n=1}^{N} \eta_n G^2 \;&\leqslant\; \frac{G^2}{2L_1} \sum_{n=1}^{N} \frac{1}{n} \leqslant \frac{G^2}{2L_1} \left(1 + \int_1^N \frac{1}{t} dt \right) \\
&=\; \frac{G^2}{2L_1}(1 + \log N).
\end{aligned}$$

Therefore, we get a loosened bound

$$\mathcal{R}_N \geq -\frac{G^2}{4L_1 C_k} \left(\frac{1 + \log N}{N} \right).$$

□

References

[1] P. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Annual Conference on Neural Information Processing Systems*, 2007.