AGGREGATED EXPECTILE REGRESSION BY EXPONENTIAL WEIGHTING

Yuwen Gu and Hui Zou

University of Connecticut and University of Minnesota

Supplementary Material

This supplementary file contains the proofs of the two theorems in Section 2 of the main article. More numerical studies are also included to illustrate various aspects of the aggregation algorithms.

S1. Proofs

We present here the proofs of all theoretical results in previous sections along with a few technical lemmas. The first lemma concerns the smoothness of the asymmetric squared error loss.

Lemma 1. *The asymmetric squared error loss Ψ_τ has Lipschitz continuous derivative, that is,*

$$2\underline{c}|u-u_0| \le |\Psi'_\tau(u) - \Psi'_\tau(u_0)| \le 2\bar{c}|u-u_0|,\ \forall u, u_0 \in \mathbb{R}. \tag{S1.1}$$

Moreover, Ψ_τ also satisfies

$$\begin{aligned}\underline{c}(u-u_0)^2 &\le \Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u-u_0)\\ &\le \bar{c}(u-u_0)^2,\ \forall u, u_0 \in \mathbb{R}.\end{aligned} \tag{S1.2}$$

Proof. We first prove the inequalities in (S1.1). For ease of notation, let $w_\tau(u) = |\tau - I(u < 0)|$. Observe that $\underline{c} \le w_\tau(u) \le \bar{c}$ for all $u \in \mathbb{R}$. Note that $\Psi'_\tau(u) = 2w_\tau(u)u$. If $u = 0$ or $u_0 = 0$, then the inequalities in (S1.1) hold trivially. If $uu_0 > 0$, we must have $w_\tau(u) = w_\tau(u_0)$. It follows that

$$2\underline{c}|u-u_0| \le |\Psi'_\tau(u) - \Psi'_\tau(u_0)| = 2w_\tau(u)|u-u_0| \le 2\bar{c}|u-u_0|.$$

If instead $uu_0 < 0$, by the symmetric roles of u and u_0, we can assume without loss of generality that $u > 0$ and $u_0 < 0$. It follows that

$$2\underline{c}|u-u_0| \le |\Psi'_\tau(u) - \Psi'_\tau(u_0)| = 2\tau u - 2(1-\tau)u_0 \le 2\bar{c}|u-u_0|.$$

This establishes the inequalities in (S1.1).

Next we prove the inequalities in (S1.2). Note that the second inequality in (S1.2) follows from the second inequality in (S1.1) by Theorem 2.1.5

of Nesterov (2004). To prove the first inequality in (S1.2), note that

$$\begin{aligned}
&\Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u-u_0) \\
=& \, w_\tau(u)u^2 - w_\tau(u_0)u_0^2 - 2w_\tau(u_0)u_0(u-u_0) \\
=& \, w_\tau(u_0)(u-u_0)^2 + \{w_\tau(u) - w_\tau(u_0)\}u^2.
\end{aligned}$$

If $w_\tau(u) \geq w_\tau(u_0)$, then obviously we get

$$\Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u-u_0) \geq w_\tau(u_0)(u-u_0)^2 \geq \underline{c}(u-u_0)^2.$$

If $w_\tau(u) < w_\tau(u_0)$, then we have $\underline{c} = w_\tau(u)$, $\bar{c} = w_\tau(u_0)$ and $u_0 u \leq 0$. It follows that

$$\begin{aligned}
&\Psi_\tau(u) - \Psi_\tau(u_0) - \Psi'_\tau(u_0)(u-u_0) \\
=& \, \underline{c}u^2 - 2\bar{c}u_0(u-u_0) - \bar{c}u_0^2 \\
\geq& \, \underline{c}u^2 - 2\underline{c}u_0 u + \underline{c}u_0^2 = \underline{c}(u-u_0)^2.
\end{aligned}$$

Therefore, we have established the first inequality in (S1.2). This completes the proof of Lemma 1. □

The second lemma explores some properties of sub-exponential random variables. See Vershynin (2010) for a thorough treatment of this family of random variables.

Lemma 2. *Let ζ be a centered sub-exponential random variable, whose sub-exponential norm satisfies $K = \|\zeta\|_{SEXP} = \sup_{k\geq 1} k^{-1}(\mathbb{E}|\zeta|^k)^{1/k} \in (0, \infty)$. Then, the following two results hold:*

(a). $\mathbb{E}\exp(t|\zeta|) \leq 2\exp(CK^2t^2)$, $\forall |t| \leq c/K$, *where* $C = 2e^2, c = 1/(2e)$ *and* $e = \exp(1)$.

(b). Let $\eta_\tau = \Psi'_\tau(\zeta - \mathscr{E}^\tau(\zeta)) = 2(\zeta - \mathscr{E}^\tau(\zeta))|\tau - I(\zeta < \mathscr{E}^\tau(\zeta))|$ *for* $\tau \in (0, 1)$. *Then* η_τ *is also centered and satisfies*

$$\mathbb{E}\exp(t|\eta_\tau|) \leq 2\exp(CK_\tau^2t^2),\ \forall |t| \leq c/K_\tau,$$

and

$$\mathbb{E}\{|\eta_\tau|^2\exp(t|\eta_\tau|)\} \leq 16\sqrt{2}K_\tau^2\exp(2C^2K_\tau^2t^2),\ \forall |t| \leq c/(2K_\tau),$$

where $K_\tau = \|\eta_\tau\|_{SEXP} = \sup_{k\geq 1} k^{-1}(\mathbb{E}|\eta_\tau|^k)^{1/k}$ *is the sub-exponential norm of* η_τ *satisfying that* $K_\tau \leq 2\bar{c}\{K + |\mathscr{E}^\tau(\zeta)|\}$.

Proof. Let us first show result (a). It follows directly from Lemma 5.15 of Vershynin (2010) that $\mathbb{E}\exp(t\zeta) \leq \exp(CK^2t^2)$, $\forall |t| \leq c/K$. Let F be the CDF of ζ. For $|t_0| \leq c/K$ and $t_0 \geq 0$, we have $\mathbb{E}\exp(t_0\zeta) \leq \exp(CK^2t_0^2)$

and $\mathbb{E}\exp(-t_0\zeta) \leq \exp(CK^2t_0^2)$. It then follows that

$$\int_0^\infty \exp(t_0 z)\,\mathrm{d}F(z) \leq \exp(CK^2t_0^2), \quad \text{and}$$
$$\int_{-\infty}^0 \exp(-t_0 z)\,\mathrm{d}F(z) \leq \exp(CK^2t_0^2).$$

Thus, we have

$$\begin{aligned}\mathbb{E}\exp(t_0|\zeta|) &= \int_0^\infty \exp(t_0 z)\,\mathrm{d}F(z) + \int_{-\infty}^0 \exp(-t_0 z)\,\mathrm{d}F(z)\\ &\leq 2\exp(CK^2t_0^2).\end{aligned}$$

Now for any $t \in [-c/K, c/K]$, we have $\mathbb{E}\exp(t|\zeta|) \leq \mathbb{E}\exp(|t|\cdot|\zeta|) \leq 2\exp(CK^2t^2)$. This completes the proof of result (a).

For result (b), first note that by definition of $\mathscr{E}^\tau(\zeta)$, we conclude that $\mathbb{E}(\eta_\tau) = 0$. By Minkowski inequality, we have $K_\tau \leq 2\bar{c}\{K + |\mathscr{E}^\tau(\zeta)|\} < \infty$. Thus, η_τ is also a sub-exponential random variable. The upper bound on the moment generating function of $|\eta_\tau|$ follows naturally from result (a). For $\mathbb{E}\{|\eta_\tau|^2\exp(t|\eta_\tau|)\}$, note that by Cauchy-Schwarz inequality we have

$$\mathbb{E}\{|\eta_\tau|^2\exp(t|\eta_\tau|)\} \leq (\mathbb{E}|\eta_\tau|^4)^{1/2}\{\mathbb{E}\exp(2t|\eta_\tau|)\}^{1/2},$$

for which $(\mathbb{E}|\eta_\tau|^4)^{1/2} = \{(\mathbb{E}|\eta_\tau|^4)^{1/4}\}^2 \leq (4K_\tau)^2$ and $\{\mathbb{E}\exp(2t|\eta_\tau|)\}^{1/2} \leq \sqrt{2}\exp(2CK_\tau^2t^2)$ for any $|t| \leq c/(2K_\tau)$. Result (b) then follows. □

Proof of Theorem 1. We first prove the oracle inequality for AEREW by Algorithm 1. The same proof works for AEREW by Algorithm 2 with slight modification which will be explained later.

Let $q_{n_0}^n = \sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \sum_{i=n_0+1}^n \Psi_\tau(y_i - \hat{e}_{\tau,\, j,\, n_0}(\mathbf{x}_i))\right\}$. Observe that

$$\begin{aligned} q_{n_0}^n =& \sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \Psi_\tau(y_{n_0+1} - \hat{e}_{\tau,\, j,\, n_0}(\mathbf{x}_{n_0+1}))\right\} \\ & \times \frac{\sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \sum_{i=n_0+1}^{n_0+2} \Psi_\tau(y_i - \hat{e}_{\tau,\, j,\, n_0}(\mathbf{x}_i))\right\}}{\sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \Psi_\tau(y_{n_0+1} - \hat{e}_{\tau,\, j,\, n_0}(\mathbf{x}_{n_0+1}))\right\}} \\ & \times \cdots \times \frac{\sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \sum_{i=n_0+1}^{n} \Psi_\tau(y_i - \hat{e}_{\tau,\, j,\, n_0}(\mathbf{x}_i))\right\}}{\sum_{j=1}^\infty \pi_j \exp\left\{-\lambda \sum_{i=n_0+1}^{n-1} \Psi_\tau(y_i - \hat{e}_{\tau,\, j,\, n_0}(\mathbf{x}_i))\right\}} \\ =& \prod_{i=n_0+1}^n \left(\sum_{j=1}^\infty W_{j,\, i} \exp\left\{-\lambda \Psi_\tau(y_i - \hat{e}_{\tau,\, j,\, n_0}(\mathbf{x}_i))\right\} \right). \end{aligned}$$

Fix $i \in \{n_0+1, \dots, n\}$. Let J be the discrete random variable such that $\mathbb{P}(J = j) = W_{j,i}$, $j \geq 1$. Let ν be the discrete measure induced by J on $\mathbb{Z}^+$ such that $\nu(j) = \mathbb{P}(J = j) = W_{j,i}$, $j \geq 1$. For ease of notation, denote $h(J) = -\Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))$. It follows that

$$\sum_{j=1}^\infty W_{j,i} \exp\{-\lambda \Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))\} = \mathbb{E}_\nu \exp\{\lambda h(J)\}.$$

By Lemma 3.6.1 of Catoni and Picard (2004, p. 85), we have

$$\begin{aligned}\log \mathbb{E}_\nu \exp\{\lambda h(J)\} \leq \lambda \mathbb{E}_\nu(h(J)) + \frac{\lambda^2}{2}\mathrm{var}_\nu(h(J)) \\ \cdot \exp\left[\lambda \max\left\{0, \sup_{\gamma\in[0,\lambda]} \frac{\boldsymbol{M}^3_{\nu_\gamma}(h(J))}{\mathrm{var}_{\nu_\gamma}(h(J))}\right\}\right],\end{aligned} \tag{S1.3}$$

where the induced measure $\nu_\gamma, \gamma \in [0, \lambda]$ is given by

$$\nu_\gamma(j) = \frac{W_{j,\,i}\exp(\gamma h(j))}{\sum_{j'=1}^{\infty} W_{j',\,i}\exp(\gamma h(j'))},\ j \geq 1,$$

and $\boldsymbol{M}^3_{\nu_\gamma}(h(J)) = \mathbb{E}_{\nu_\gamma}\{h(J) - \mathbb{E}_{\nu_\gamma} h(J)\}^3$ is the third central moment.

To facilitate the presentation, let $b_\tau(\mathbf{x}) = \mathscr{E}^\tau(\varepsilon|\mathbf{x})$ be the τth conditional expectile of the random error ε given $\mathbf{X} = \mathbf{x}$. It can be seen that the τth conditional expectile function of Y given $\mathbf{X} = \mathbf{x}$ is $e_\tau(\mathbf{x}) = m(\mathbf{x}) + \sigma(\mathbf{x})b_\tau(\mathbf{x})$. By Lemma 1, it can be seen that

$$\begin{aligned}\sup_{\gamma\in[0,\lambda]} \frac{\boldsymbol{M}^3_{\nu_\gamma}(h(J))}{\mathrm{var}_{\nu_\gamma}(h(J))} &\leq \sup_{\gamma\in[0,\lambda]} \sup_{j\geq 1} |h(j) - \mathbb{E}_{\nu_\gamma}(h(J))| \leq \sup_{j_1,\,j_2\geq 1} |h(j_1) - h(j_2)| \\ &\leq 2\sup_{j\geq 1} |\Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - e_\tau(\mathbf{x}_i))| \\ &\leq 2\sigma(\mathbf{x}_i)|\Psi'_\tau(\varepsilon_i - b_\tau(\mathbf{x}))| \sup_{j\geq 1} |\hat{e}_{\tau,\,j,\,n_0}(\mathbf{x}_i) - e_\tau(\mathbf{x}_i)| \\ &\quad + 2\bar{c}\sup_{j\geq 1}(\hat{e}_{\tau,j,n_0}(\mathbf{x}_i) - e_\tau(\mathbf{x}_i))^2\end{aligned}$$

and that

$$\begin{aligned}
&\mathrm{var}_\nu(h(J)) \le \mathbb{E}_\nu\big\{\Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\big\}^2\\
&\le \sup_{j\ge 1}\big(|\Psi'_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))| + \bar{c}|\hat{e}_{\tau,j,n_0}(\mathbf{x}_i) - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i)|\big)^2\\
&\qquad\cdot \mathbb{E}_\nu\big(\hat{e}_{\tau,J,n_0}(\mathbf{x}_i) - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i)\big)^2\\
&\le \Big\{\sigma(\mathbf{x}_i)|\Psi'_\tau(\varepsilon_i - b_\tau(\mathbf{x}))| + 4\bar{c}\sup_{j\ge 1}|\hat{e}_{\tau,j,n_0}(\mathbf{x}_i) - e_\tau(\mathbf{x}_i)|\Big\}^2\\
&\qquad\cdot \mathbb{E}_\nu\big(\hat{e}_{\tau,J,n_0}(\mathbf{x}_i) - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i)\big)^2.
\end{aligned}$$

Also from Lemma 1, we get that

$$\begin{aligned}
&\Psi_\tau(y_i - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\\
&\ge \Psi'_\tau(y_i - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))(\mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i) - \hat{e}_{\tau,j,n_0}(\mathbf{x}_i))\\
&\qquad + \underline{c}(\hat{e}_{\tau,j,n_0}(\mathbf{x}_i) - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))^2.
\end{aligned}$$

Taking expectation with respect to J on both sides of the above inequality, we have

$$\begin{aligned}
&\mathbb{E}_\nu(\hat{e}_{\tau,J,n_0}(\mathbf{x}_i) - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))^2\\
&\le \underline{c}^{-1}\big\{\mathbb{E}_\nu\Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - \mathbb{E}_\nu\hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\big\}.
\end{aligned}$$

Let $\xi_i = \Psi'_\tau(\varepsilon_i - b_\tau(\mathbf{x}_i))$. It follows from inequality (S1.3) and assumptions

(C1) – (C3) that with probability one

$$\begin{aligned}
&\log \mathbb{E}_\nu \exp\{\lambda h(J)\} \\
&\leq \lambda \mathbb{E}_\nu(h(J)) + \frac{\lambda^2}{2}(C_0|\xi_i| + 4\bar{c}A_\tau)^2 \exp\left\{2\lambda C_0 A_\tau |\xi_i| + 2\lambda \bar{c} A_\tau^2\right\} \\
&\qquad \cdot \underline{c}^{-1}\left\{\mathbb{E}_\nu \Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) - \Psi_\tau(y_i - \mathbb{E}_\nu \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\right\} \\
&\leq -\lambda \mathbb{E}_\nu \Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) + \frac{\lambda^2}{\underline{c}} \exp(2\lambda\bar{c}A_\tau^2)\left(C_0^2|\xi_i|^2 + 16\bar{c}^2 A_\tau^2\right) \\
&\qquad \cdot \exp(2\lambda C_0 A_\tau |\xi_i|)\left\{\mathbb{E}_\nu \Psi_\tau(y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\right. \\
&\qquad\qquad\qquad\qquad \left. - \Psi_\tau(y_i - \mathbb{E}_\nu \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\right\}.
\end{aligned} \tag{S1.4}$$

Take the expectation (denoted by $\mathbb{E}_i$) of both sides of (S1.4) with respect to Y_i conditional on $\mathbf{x}_i \cup (y_k, \mathbf{x}_k)_{k=1}^{i-1}$. By Lemma 2, when λ is chosen small enough such that $2\lambda C_0 A_\tau \leq (4eK_\tau)^{-1}$, with probability one we have

$$\begin{aligned}
&\mathbb{E}_i \log\left(\mathbb{E}_\nu \exp\left\{-\lambda \Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\right\}\right) \\
&\leq -\lambda \mathbb{E}_i\left\{\mathbb{E}_\nu \Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\right\} \\
&+ \lambda^2 \underline{c}^{-1} \exp(2\lambda\bar{c}A_\tau^2)\left\{C_0^2 \mathscr{M}_2(2\lambda C_0 A_\tau) + 16\bar{c}^2 A_\tau \mathscr{M}_0(2\lambda C_0 A_\tau)\right\} \\
&\times \mathbb{E}_i\left[\mathbb{E}_\nu \Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) - \Psi_\tau(Y_i - \mathbb{E}_\nu \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\right].
\end{aligned}$$

Moreover, if λ also satisfies

$$\lambda^2 \underline{c}^{-1} \exp(2\lambda\bar{c}A_\tau^2)\left\{C_0^2 \mathscr{M}_2(2\lambda C_0 A_\tau) + 16\bar{c}^2 A_\tau \mathscr{M}_0(2\lambda C_0 A_\tau)\right\} \leq \lambda,$$

with probability one we will have

$$\mathbb{E}_i \log\left(\mathbb{E}_\nu \exp\left\{-\lambda \Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i))\right\}\right) \le -\lambda \mathbb{E}_i \Psi_\tau(Y_i - \mathbb{E}_\nu \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)),$$

since by convexity of $\Psi_\tau(\cdot)$ and Jensen's inequality we have

$$\Psi_\tau(Y_i - \mathbb{E}_\nu \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)) \le \mathbb{E}_\nu \Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{x}_i)).$$

It follows that when λ is small enough such that condition (2.2) holds, we have

$$\begin{aligned}
\mathbb{E}\log(1/q_{n_0}^n) &= -\sum_{i=n_0+1}^{n} \mathbb{E}\log\left(\sum_{j=1}^{\infty} W_{j,i} \exp\left\{-\lambda \Psi_\tau(Y_i - \hat{e}_{\tau,j,n_0}(\mathbf{X}_i))\right\}\right) \\
&= -\sum_{i=n_0+1}^{n} \mathbb{E}\Big[\mathbb{E}_i \log\left(\mathbb{E}_\nu \exp\left\{-\lambda \Psi_\tau(Y_i - \hat{e}_{\tau,J,n_0}(\mathbf{X}_i))\right\}\right)\Big] \\
&\ge \lambda \mathbb{E}\Bigg[\sum_{i=n_0+1}^{n} \mathbb{E}_i \Psi_\tau\Bigg(Y_i - \sum_{j=1}^{\infty} W_{j,i} \hat{e}_{\tau,j,n_0}(\mathbf{X}_i)\Bigg)\Bigg] \\
&= \lambda \sum_{i=n_0+1}^{n} \mathbb{E}\Psi_\tau\Bigg(Y - \sum_{j=1}^{\infty} W_{j,\,i} \hat{e}_{\tau,j,n_0}(\mathbf{X})\Bigg).
\end{aligned}$$

The last equality is due to the independence of the observations, i.e., $(Y, \mathbf{X})$ is independent of $(Y_i, \mathbf{X}_i)_{i=1}^n$. On the other hand, we have, for each $j^* \ge 1$,

$$\begin{aligned}
\mathbb{E}\log(1/q_{n_0}^n) &\le \log(1/\pi_{j^*}) + \lambda \sum_{i=n_0+1}^{n} \mathbb{E}\Psi_\tau(Y_i - \hat{e}_{\tau,j^*,n_0}(\mathbf{X}_i)) \\
&= \log(1/\pi_{j^*}) + \lambda(n - n_0)\mathbb{E}\Psi_\tau(Y - \hat{e}_{\tau,j^*,n_0}(\mathbf{X})).
\end{aligned}$$

Therefore, for any $j^* \geq 1$, we have

$$\begin{aligned}&\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\mathbb{E}\Psi_\tau\left(Y-\sum_{j=1}^{\infty}W_{j,i}\hat{e}_{\tau,j,n_0}(\mathbf{X})\right)\\&\leq\frac{\log(1/\pi_{j^*})}{\lambda(n-n_0)}+\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,j^*,n_0}(\mathbf{X})).\end{aligned}\tag{S1.5}$$

Note that by definition of $\hat{e}_{\tau,\cdot,n}(\mathbf{x})$, we have

$$y-\hat{e}_{\tau,\cdot,n}(\mathbf{x})=\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\left(y-\sum_{j=1}^{\infty}W_{j,i}\hat{e}_{\tau,j,n_0}(\mathbf{x})\right).$$

It follows from (S1.5) and convexity of $\Psi_\tau(\cdot)$ that for each $j^* \geq 1$,

$$\begin{aligned}\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,\cdot,n}(\mathbf{X}))&\leq\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\mathbb{E}\Psi_\tau\left(Y-\sum_{j=1}^{\infty}W_{j,i}\hat{e}_{\tau,j,n_0}(\mathbf{X})\right)\\&\leq\frac{\log(1/\pi_{j^*})}{\lambda(n-n_0)}+\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,j^*,n_0}(\mathbf{X})).\end{aligned}$$

This completes the proof of inequality (2.3). To show (2.4), note that by Lemma 1

$$\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,j^*,n_0}(\mathbf{X}))\leq\mathbb{E}\Psi_\tau(Y-e_\tau(\mathbf{X}))+\bar{c}\mathbb{E}(e_\tau(\mathbf{X})-\hat{e}_{\tau,j^*,n_0}(\mathbf{X}))^2$$

$$\mathbb{E}\Psi_\tau(Y-\hat{e}_{\tau,\cdot,n}(\mathbf{X}))\geq\mathbb{E}\Psi_\tau(Y-e_\tau(\mathbf{X}))+\underline{c}\mathbb{E}(e_\tau(\mathbf{X})-\hat{e}_{\tau,\cdot,n}(\mathbf{X}))^2$$

due to the fact that $\mathbb{E}\{\Psi_\tau'(Y-e_\tau(\mathbf{X}))|\mathbf{X}\}=0$. Inequality (2.4) then follows from (2.3).

To prove the same result for AEREW by Algorithm 2, we note by convexity of $\Psi_\tau(\cdot)$ that

$$\Psi_\tau\big(y - \hat{e}^B_{\tau,\cdot,n}(\mathbf{x})\big) \le \frac{1}{B}\sum_{k=1}^{B}\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\Psi_\tau\left(y - \sum_{j=1}^{\infty} W^{(k)}_{j,i}\hat{e}^{(k)}_{\tau,j,n_0}(\mathbf{x})\right).$$

The result then follows from the previous proof for AEREW by Algorithm 1.

□

Proof of Theorem 2. The proof is similar to that of Theorem 1 with slight modifications. Define $q^n_{n_0} = \sum_{j=1}^{\infty}\pi_j \exp\left\{-\lambda\sum_{i=n_0+1}^{n}\Psi_\tau(y_i - \hat{e}_{\tau,j,i})\right\}$. It can be shown that

$$q^n_{n_0} = \prod_{i=n_0+1}^{n}\left(\sum_{j=1}^{\infty}\Lambda_{j,i}\exp\left\{-\lambda\Psi_\tau(y_i - \hat{e}_{\tau,j,i})\right\}\right).$$

For each $i = n_0+1,\ldots,n$, let J^i be the discrete random variable such that $\mathbb{P}(J^i = j) = \Lambda_{j,i}$, $j \ge 1$. Let ν^i be the discrete measure induced by J^i on $\mathbb{Z}^+$ such that $\nu^i(j) = \mathbb{P}(J^i = j) = \Lambda_{j,i}$, $j \ge 1$. For ease of notation, denote $h(J^i) = -\Psi_\tau(y_i - \hat{e}_{\tau,J^i,i})$. It follows that

$$\sum_{j=1}^{\infty}\Lambda_{j,i}\exp\{-\lambda\Psi_\tau(y_i - \hat{e}_{\tau,j,i})\} = \mathbb{E}_{\nu^i}\exp\{\lambda h(J^i)\}.$$

By Lemma 3.6.1 of Catoni and Picard (2004, p. 85), we have

$$\begin{aligned}\log \mathbb{E}_{\nu^i} \exp\{\lambda h(J^i)\} \leq \lambda \mathbb{E}_{\nu^i}(h(J^i)) + \frac{\lambda^2}{2}\mathrm{var}_{\nu^i}(h(J^i)) \\ \cdot \exp\left[\lambda \max\left\{0, \sup_{\gamma\in[0,\lambda]} \frac{\boldsymbol{M}^3_{\nu^i_\gamma}(h(J^i))}{\mathrm{var}_{\nu^i_\gamma}(h(J^i))}\right\}\right],\end{aligned} \tag{S1.6}$$

where the induced measure $\nu^i_\gamma, \gamma \in [0, \lambda]$ is given by

$$\nu^i_\gamma(j) = \frac{\Lambda_{j,\,i}\exp(\gamma h(j))}{\sum_{j'=1}^{\infty}\Lambda_{j',\,i}\exp(\gamma h(j'))},\ j \geq 1,$$

and $\boldsymbol{M}^3_{\nu^i_\gamma}(h(J^i)) = \mathbb{E}_{\nu^i_\gamma}\{h(J^i) - \mathbb{E}_{\nu^i_\gamma} h(J^i)\}^3$ is the third central moment.

Note that the τth conditional expectile function of Y given $\mathbf{X}_i = \mathbf{x}_i$ and $Z^{i-1} = z^{i-1}$ can be expressed as $e_{\tau,i} = m_i + \sigma_i b_{\tau,i}$, where $b_{\tau,i} = \mathcal{E}^\tau(\varepsilon_i|\mathbf{X}_i = \mathbf{x}_i, Z^{i-1} = z^{i-1})$. By Lemma 1, it can be shown that

$$\begin{aligned}\sup_{\gamma\in[0,\lambda]} \frac{\boldsymbol{M}^3_{\nu^i_\gamma}(h(J^i))}{\mathrm{var}_{\nu^i_\gamma}(h(J^i))} &\leq \sup_{\gamma\in[0,\lambda]}\sup_{j\geq1}|h(j) - \mathbb{E}_{\nu^i_\gamma}(h(J^i))| \leq \sup_{j_1,\,j_2\geq1}|h(j_1) - h(j_2)| \\ &\leq 2\sup_{j\geq1}|\Psi_\tau(y_i - \hat{e}_{\tau,j,i}) - \Psi_\tau(y_i - e_{\tau,i})| \\ &\leq 2\sigma_i|\Psi'_\tau(\varepsilon_i - b_{\tau,i})|\sup_{j\geq1}|\hat{e}_{\tau,\,j,\,i} - e_{\tau,i}| \\ &\quad + 2\bar{c}\sup_{j\geq1}(\hat{e}_{\tau,j,i} - e_{\tau,i})^2\end{aligned}$$

and that

$$\begin{aligned}
&\operatorname{var}_{\nu^i}(h(J^i)) \le \mathbb{E}_{\nu^i}\big\{\Psi_\tau(y_i-\hat{e}_{\tau,J^i,i})-\Psi_\tau(y_i-\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i})\big\}^2\\
&\le \sup_{j\ge 1}\big(|\Psi_\tau'(y_i-\hat{e}_{\tau,j,i})|+\bar{c}|\hat{e}_{\tau,j,i}-\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i}|\big)^2\mathbb{E}_{\nu^i}\big(\hat{e}_{\tau,J^i,i}-\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i}\big)^2\\
&\le \Big\{\sigma_i|\Psi_\tau'(\varepsilon_i-b_{\tau,i})|+4\bar{c}\sup_{j\ge 1}|\hat{e}_{\tau,j,i}-e_{\tau,i}|\Big\}^2\mathbb{E}_{\nu^i}\big(\hat{e}_{\tau,J^i,i}-\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i}\big)^2.
\end{aligned}$$

Also from Lemma 1, we get that

$$\begin{aligned}
&\Psi_\tau(y_i-\hat{e}_{\tau,j,i})-\Psi_\tau(y_i-\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i})\\
&\ge \Psi_\tau'(y_i-\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i})(\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i}-\hat{e}_{\tau,j,i})+\underline{c}(\hat{e}_{\tau,j,i}-\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i})^2.
\end{aligned}$$

Taking expectation with respect to J^i on both sides of the above inequality, we have

$$\mathbb{E}_{\nu^i}(\hat{e}_{\tau,J^i,i}-\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i})^2 \le \underline{c}^{-1}\big\{\mathbb{E}_{\nu^i}\Psi_\tau(y_i-\hat{e}_{\tau,J^i,i})-\Psi_\tau(y_i-\mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i})\big\}.$$

Let $\xi_i=\Psi_\tau'(\varepsilon_i-b_{\tau,i})$. It follows from inequality (S1.6) and assumptions

(C1') – (C3') that with probability one

$$\begin{aligned}
&\log \mathbb{E}_{\nu^i} \exp\big\{\lambda h(J^i)\big\}\\
&\leq \lambda \mathbb{E}_{\nu^i}(h(J^i)) + \frac{\lambda^2}{2}(C_0|\xi_i| + 4\bar{c}A_\tau)^2 \exp\big\{2\lambda C_0 A_\tau |\xi_i| + 2\lambda \bar{c} A_\tau^2\big\}\\
&\qquad \cdot \underline{c}^{-1}\big\{\mathbb{E}_{\nu^i}\Psi_\tau(y_i - \hat{e}_{\tau,J,i}) - \Psi_\tau(y_i - \mathbb{E}_\nu \hat{e}_{\tau,J,i})\big\} \qquad \text{(S1.7)}\\
&\leq -\lambda \mathbb{E}_{\nu^i}\Psi_\tau(y_i - \hat{e}_{\tau,J^i,i}) + \lambda^2 \underline{c}^{-1}\exp(2\lambda\bar{c}A_\tau^2)\left(C_0^2|\xi_i|^2 + 16\bar{c}^2 A_\tau^2\right)\\
&\qquad \cdot \exp(2\lambda C_0 A_\tau|\xi_i|)\big\{\mathbb{E}_{\nu^i}\Psi_\tau(y_i - \hat{e}_{\tau,J,i}) - \Psi_\tau(y_i - \mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i})\big\}.
\end{aligned}$$

Take the expectation (denoted by $\mathbb{E}_i$) of both sides of (S1.7) with respect to Y_i conditional on $\mathbf{X}_i = \mathbf{x}_i$ and $Z^{i-1} = z^{i-1}$. Note that when λ satisfies (2.2), we have $2\lambda C_0 A_\tau \leq (4eK_\tau)^{-1}$. By Lemma 2, with probability one we get

$$\begin{aligned}
&\mathbb{E}_i \log\left(\mathbb{E}_{\nu^i}\exp\big\{-\lambda\Psi_\tau(Y_i - \hat{e}_{\tau,J^i,i})\big\}\right)\\
&\leq -\lambda\mathbb{E}_i\big\{\mathbb{E}_{\nu^i}\Psi_\tau(Y_i - \hat{e}_{\tau,J^i,i})\big\}\\
&+ \lambda^2\underline{c}^{-1}\exp(2\lambda\bar{c}A_\tau^2)\left\{C_0^2\mathscr{M}_2(2\lambda C_0 A_\tau) + 16\bar{c}^2 A_\tau \mathscr{M}_0(2\lambda C_0 A_\tau)\right\}\\
&\times \mathbb{E}_i\big[\mathbb{E}_{\nu^i}\Psi_\tau(Y_i - \hat{e}_{\tau,J^i,i}) - \Psi_\tau(Y_i - \mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i})\big].
\end{aligned}$$

Moreover, inequality (2.2) also implies that

$$\lambda^2\underline{c}^{-1}\exp(2\lambda\bar{c}A_\tau^2)\left\{C_0^2\mathscr{M}_2(2\lambda C_0 A_\tau) + 16\bar{c}^2 A_\tau\mathscr{M}_0(2\lambda C_0 A_\tau)\right\} \leq \lambda,$$

then with probability one we will have

$$\mathbb{E}_i \log\big(\mathbb{E}_{\nu^i} \exp\big\{-\lambda\Psi_\tau(Y_i - \hat{e}_{\tau,J^i,i})\big\}\big) \le -\lambda\mathbb{E}_i\Psi_\tau(Y_i - \mathbb{E}_{\nu^i}\hat{e}_{\tau,J^i,i}),$$

since by convexity of $\Psi_\tau(\cdot)$ and Jensen's inequality we have

$$\Psi_\tau(Y_i - \mathbb{E}_{\nu^i}\hat{e}_{\tau,J,i}) \le \mathbb{E}_{\nu^i}\Psi_\tau(Y_i - \hat{e}_{\tau,J,i}).$$

It follows that when λ satisfies inequality (2.2), we have

$$\begin{aligned}\mathbb{E}\log(1/q_{n_0}^n) &= -\sum_{i=n_0+1}^{n}\mathbb{E}\log\left(\sum_{j=1}^{\infty}\Lambda_{j,i}\exp\left\{-\lambda\Psi_\tau(Y_i - \hat{e}_{\tau,j,i})\right\}\right)\\ &= -\sum_{i=n_0+1}^{n}\mathbb{E}\big[\mathbb{E}_i\log\big(\mathbb{E}_{\nu^i}\exp\big\{-\lambda\Psi_\tau(Y_i - \hat{e}_{\tau,J^i,i})\big\}\big)\,\big]\\ &\ge \lambda\mathbb{E}\left[\sum_{i=n_0+1}^{n}\mathbb{E}_i\Psi_\tau\left(Y_i - \sum_{j=1}^{\infty}\Lambda_{j,i}\hat{e}_{\tau,j,i}\right)\right].\end{aligned}$$

On the other hand, we have, for each $j^* \ge 1$,

$$\mathbb{E}\log(1/q_{n_0}^n) \le \log(1/\pi_{j^*}) + \lambda\sum_{i=n_0+1}^{n}\mathbb{E}\Psi_\tau(Y_i - \hat{e}_{\tau,j^*,i}).$$

Therefore, for any $j^* \geq 1$, we have

$$\begin{aligned}&\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\mathbb{E}\Psi_\tau(Y_i-\hat{e}_{\tau,\cdot,i})\\&\leq\frac{\log(1/\pi_{j^*})}{\lambda(n-n_0)}+\frac{1}{n-n_0}\sum_{i=n_0+1}^{n}\mathbb{E}\Psi_\tau(Y_i-\hat{e}_{\tau,j^*,i}).\end{aligned}$$

This completes the proof of inequality (2.5). Inequality (2.6) follows from Lemma 1 and inequality (2.5) by noting that

$$\begin{aligned}\mathbb{E}\Psi_\tau(Y_i-\hat{e}_{\tau,j^*,i})&\leq\mathbb{E}\Psi_\tau(Y_i-e_{\tau,i})+\bar{c}\mathbb{E}(e_{\tau,i}-\hat{e}_{\tau,j^*,i})^2\\\mathbb{E}\Psi_\tau(Y_i-\hat{e}_{\tau,\cdot,i})&\geq\mathbb{E}\Psi_\tau(Y_i-e_{\tau,i})+\underline{c}\mathbb{E}(e_{\tau,i}-\hat{e}_{\tau,\cdot,i})^2\end{aligned}$$

due to the fact that $\mathbb{E}_i\{\Psi'_\tau(Y_i-e_{\tau,i})|\mathbf{X}_i=\mathbf{x}_i,Z^{i-1}=z^{i-1}\}=0$. This completes the proof. □

S2. Effect of including a biased candidate model on aggregation

In this section, we investigate the effect of including an obviously biased candidate model on the performance of the aggregated procedure. Recall that in model (3.2), we applied AEREW to aggregate the local linear expectile regressions with different bandwidths. Based on the design of the model, we know that the multiple linear expectile regression will produce biased estimates. Complement to the numerical study in Section 3.1, we carry out further numerical study by also including the multiple linear expectile regression as a candidate model. The results are summarized in Table S.1. First, according to the design of our model, the linear expectile regression becomes more and more biased as τ grows. The p_{CV} from Table S.1 justifies this since the linear expectile regression is selected less and less often by the cross-validation as τ grows. Also from Table S.1, we can see that adding the linear expectile regression model indeed has deteriorated the performance of the aggregated procedure, but the impact is small since AEREW can adaptively select weights that favor procedures with good performance.

Similarly, we note that in Table 4, HS100 is the worst method in terms of prediction risk and we can also gain some insights from investigating the effect of removing HS100 on the aggregated procedure. We report in Table S.2 the performance of the aggregated procedure by combining only linear and boosted expectile regressions with lag 20. Compared to Table 4, we can see that indeed the performance of the aggregated procedure has improved.

In practice, if there are obviously very bad models (in terms of prediction), we recommend ruling those out before applying the aggregation.

Table S.1: Estimated prediction risks and MSDs of local linear (LL) regressions with five candidate bandwidths, the linear expectile regression (LM), the five-fold cross-validated kernel estimator, and AEREW ($\lambda = 1$) for the heteroscedastic model in simulation 1. The numbers listed are averages over 100 independent runs with their respective standard errors reported in the parentheses. The proportion of each candidate estimator being selected by the five-fold cross-validation among these 100 runs is reported by p_{CV}. All numbers are of order 10^{-2} except those corresponding to p_{CV}.

		LL (bandwidth h)					LM	CV	Aggregation
τ	Measures	0.0347	0.104	0.173	0.243	0.312			
0.05	Risk	3.73	2.98	2.89	2.90	2.92	2.96	2.95	2.84
		(0.04)	(0.02)	(0.02)	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)
	MSD	25.86	17.06	15.41	15.88	16.68	18.08	16.98	14.16
		(0.33)	(0.41)	(0.35)	(0.30)	(0.25)	(0.16)	(0.30)	(0.27)
	p_{CV}	0.00	0.21	0.17	0.13	0.10	0.39	–	–
0.10	Risk	4.98	4.21	4.13	4.18	4.24	4.37	4.22	4.10
		(0.05)	(0.04)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
	MSD	22.24	14.78	13.95	15.15	16.38	18.84	15.26	13.58
		(0.38)	(0.44)	(0.40)	(0.32)	(0.26)	(0.09)	(0.41)	(0.33)
	p_{CV}	0.03	0.27	0.30	0.09	0.05	0.26	–	–
0.25	Risk	6.74	5.89	5.90	6.09	6.30	7.11	5.96	5.96
		(0.06)	(0.03)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)	(0.03)
	MSD	17.79	10.72	10.92	13.16	15.24	21.91	11.47	11.67
		(0.33)	(0.22)	(0.22)	(0.17)	(0.14)	(0.05)	(0.29)	(0.22)
	p_{CV}	0.02	0.40	0.44	0.09	0.05	0.00	–	–
0.50	Risk	7.59	6.76	6.82	7.08	7.37	9.62	6.86	6.94
		(0.05)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
	MSD	16.00	9.62	10.33	12.57	14.78	25.90	10.46	11.33
		(0.28)	(0.21)	(0.20)	(0.17)	(0.15)	(0.05)	(0.26)	(0.22)
	p_{CV}	0.02	0.45	0.44	0.09	0.00	0.00	–	–
0.75	Risk	6.62	5.83	5.86	6.06	6.32	9.80	5.91	5.98
		(0.05)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
	MSD	17.08	10.01	10.27	12.33	14.59	31.14	10.71	11.43
		(0.30)	(0.23)	(0.22)	(0.20)	(0.17)	(0.10)	(0.29)	(0.25)
	p_{CV}	0.03	0.39	0.47	0.11	0.00	0.00	–	–
0.90	Risk	5.07	4.15	4.08	4.16	4.30	7.59	4.16	4.16
		(0.07)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
	MSD	22.42	14.07	12.80	14.01	15.94	39.77	13.86	13.75
		(0.38)	(0.31)	(0.30)	(0.27)	(0.23)	(0.28)	(0.36)	(0.32)
	p_{CV}	0.02	0.21	0.50	0.18	0.09	0.00	–	–
0.95	Risk	3.84	3.04	2.95	2.98	3.07	5.62	3.02	3.02
		(0.06)	(0.03)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)
	MSD	26.59	17.45	15.56	16.12	17.70	47.07	16.69	16.50
		(0.70)	(0.42)	(0.43)	(0.38)	(0.32)	(0.42)	(0.40)	(0.43)
	p_{CV}	0.00	0.24	0.31	0.24	0.21	0.00	–	–

Table S.2: Estimation risks of the linear expectile regression with lag 20, the boosted expectile regression with lag 20, and AEREW-ts ($\lambda = 0.1, 1, 10$) for the S&P 500 data.

		Individual		Aggregation		
τ	Measure	Linear	Boosting	0.1	1	10
			Series 1			
0.01	Risk	0.88	0.97	0.89	0.89	0.89
0.05	Risk	3.06	3.29	3.10	3.10	3.10
0.10	Risk	4.72	5.08	4.81	4.81	4.81
			Series 2			
0.01	Risk	0.65	0.76	0.66	0.66	0.66
0.05	Risk	1.78	1.67	1.67	1.67	1.67
0.10	Risk	2.57	2.52	2.49	2.49	2.49

S3. Split ratio for computing weights in aggregation

According to Theorem 1, the training size n_0 should be chosen such that n_0 and $n - n_0$ are of the same order as n. A typical choice is $n_0 = Cn$ for some $C \in (0, 1)$. In all of our numerical studies, we used a split ratio $C = n_0/n = 0.8$ when applying the AEREW algorithm (Algorithm 2). In this section, we investigate the impact of the split ratio on the performance of the aggregated procedure. To that end, we have run the same simulation in Section 3.1 but with training sizes $n_0 = Cn$ for $C = 0.3$ and 0.5. The results are reported in Tables S.3 and S.4. Together with Table 1, we can see that data splitting ratios $C = 0.5$ and 0.8 are slightly better than 0.3, but the differences are not very significant. In practice, we recommend using C between 0.5 and 0.8.

Table S.3: Estimated prediction risks and MSDs of local linear regressions with five candidate bandwidths, the five-fold cross-validated kernel estimator, and AEREW ($\lambda = 1$, split size $N_0 = 60$) for the heteroscedastic model in simulation 1. The numbers listed are averages over 100 independent runs with their respective standard errors reported in the parentheses. The proportion of each candidate estimator being selected by the five-fold cross-validation among these 100 runs is reported by p_{CV}. All numbers are of order 10^{-2} except those corresponding to p_{CV}.

		Bandwidth (h)					CV	Aggregation
τ	Measures	0.0347	0.104	0.173	0.243	0.312		
0.05	Risk	3.95	3.02	2.93	2.93	2.95	2.97	2.91
		(0.07)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	MSD	26.38	17.10	15.57	16.17	16.99	16.48	15.21
		(0.39)	(0.45)	(0.47)	(0.40)	(0.34)	(0.42)	(0.41)
	p_{CV}	0.01	0.31	0.23	0.11	0.34	–	–
0.10	Risk	5.03	4.13	4.07	4.13	4.20	4.12	4.07
		(0.11)	(0.03)	(0.03)	(0.03)	(0.02)	(0.03)	(0.03)
	MSD	21.96	13.89	13.08	14.45	15.88	13.99	13.17
		(0.39)	(0.35)	(0.32)	(0.28)	(0.24)	(0.34)	(0.30)
	p_{CV}	0.00	0.36	0.36	0.12	0.16	–	–
0.25	Risk	6.72	5.88	5.92	6.11	6.31	5.94	5.97
		(0.06)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	MSD	17.61	10.71	11.27	13.40	15.38	11.39	11.84
		(0.35)	(0.24)	(0.23)	(0.20)	(0.17)	(0.28)	(0.23)
	p_{CV}	0.01	0.45	0.44	0.06	0.04	–	–
0.50	Risk	7.56	6.70	6.79	7.07	7.39	6.79	6.86
		(0.05)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
	MSD	15.87	9.06	9.99	12.46	14.78	9.86	10.66
		(0.28)	(0.21)	(0.20)	(0.17)	(0.15)	(0.27)	(0.20)
	p_{CV}	0.01	0.52	0.42	0.04	0.01	–	–
0.75	Risk	6.78	5.90	5.90	6.09	6.36	5.94	5.95
		(0.06)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	MSD	17.89	10.53	10.42	12.46	14.72	10.86	10.97
		(0.32)	(0.25)	(0.24)	(0.20)	(0.16)	(0.26)	(0.24)
	p_{CV}	0.01	0.34	0.51	0.14	0.00	–	–
0.90	Risk	5.00	4.17	4.12	4.22	4.37	4.18	4.15
		(0.06)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
	MSD	21.69	13.71	12.69	14.09	16.09	13.53	12.98
		(0.32)	(0.36)	(0.36)	(0.31)	(0.26)	(0.36)	(0.35)
	p_{CV}	0.00	0.29	0.45	0.21	0.05	–	–
0.95	Risk	3.78	2.99	2.91	2.94	3.02	2.97	2.91
		(0.04)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
	MSD	25.71	16.82	14.96	15.61	17.33	15.94	14.89
		(0.35)	(0.45)	(0.44)	(0.40)	(0.35)	(0.44)	(0.42)
	p_{CV}	0.01	0.16	0.41	0.23	0.19	–	–

Table S.4: Estimated prediction risks and MSDs of local linear regressions with five candidate bandwidths, the five-fold cross-validated kernel estimator, and AEREW ($\lambda = 1$, split size $N_0 = 100$) for the heteroscedastic model in simulation 1. The numbers listed are averages over 100 independent runs with their respective standard errors reported in the parentheses. The proportion of each candidate estimator being selected by the five-fold cross-validation among these 100 runs is reported by p_{CV}. All numbers are of order 10^{-2} except those corresponding to p_{CV}.

		Bandwidth (h)					CV	Aggregation
τ	Measures	0.0347	0.104	0.173	0.243	0.312		
0.05	Risk	4.50	3.03	2.93	2.93	2.95	2.95	2.91
		(0.56)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
	MSD	26.89	17.09	15.26	15.75	16.57	16.02	14.70
		(0.72)	(0.42)	(0.41)	(0.33)	(0.26)	(0.34)	(0.35)
	p_{CV}	0.00	0.23	0.28	0.13	0.36	–	–
0.10	Risk	4.93	4.14	4.09	4.16	4.23	4.15	4.08
		(0.05)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
	MSD	21.53	13.64	12.92	14.34	15.81	13.89	12.69
		(0.33)	(0.32)	(0.29)	(0.24)	(0.21)	(0.32)	(0.27)
	p_{CV}	0.01	0.27	0.39	0.13	0.20	–	–
0.25	Risk	6.75	5.89	5.93	6.12	6.33	5.96	5.95
		(0.06)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
	MSD	17.94	10.62	11.19	13.46	15.55	11.48	11.47
		(0.37)	(0.28)	(0.25)	(0.23)	(0.20)	(0.33)	(0.26)
	p_{CV}	0.01	0.45	0.40	0.10	0.04	–	–
0.50	Risk	7.81	6.73	6.80	7.06	7.38	6.81	6.84
		(0.12)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)
	MSD	16.90	9.43	10.14	12.50	14.82	10.10	10.50
		(0.48)	(0.22)	(0.21)	(0.18)	(0.15)	(0.26)	(0.21)
	p_{CV}	0.01	0.54	0.37	0.07	0.01	–	–
0.75	Risk	6.82	5.94	5.95	6.14	6.40	5.98	5.97
		(0.07)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	MSD	17.92	10.64	10.71	12.71	14.91	11.12	10.99
		(0.41)	(0.25)	(0.22)	(0.18)	(0.15)	(0.27)	(0.23)
	p_{CV}	0.03	0.38	0.48	0.11	0.00	–	–
0.90	Risk	4.93	4.16	4.11	4.20	4.35	4.18	4.13
		(0.08)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	MSD	21.25	13.78	12.83	14.22	16.18	13.92	13.09
		(0.35)	(0.38)	(0.38)	(0.33)	(0.27)	(0.38)	(0.36)
	p_{CV}	0.00	0.29	0.41	0.24	0.06	–	–
0.95	Risk	3.92	3.06	2.96	2.99	3.08	3.04	2.97
		(0.11)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
	MSD	26.46	17.87	15.74	16.26	17.90	17.27	15.65
		(0.43)	(0.47)	(0.47)	(0.42)	(0.36)	(0.44)	(0.43)
	p_{CV}	0.00	0.32	0.25	0.23	0.20	–	–

References

Catoni, O. and Picard, J. (2004). *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, vol. 1851. Springer.

Nesterov, Y. (2004). *Introductory lectures on convex optimization*, vol. 87. Springer Science & Business Media.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027v7* .