

Stochastic optimization on complex variables and pure-state quantum tomography

Supplementary Information

A. Utreras-Alarcón,^{1,2} M. Rivera-Tapia,^{1,2} S. Niklitschek,³ and A. Delgado^{1,2}

¹*Instituto Milenio de Investigación en Óptica, Universidad de Concepción, Concepción, Chile*

²*Facultad de Ciencias Físicas y Matemáticas, Departamento de Física, Universidad de Concepción, Concepción, Chile*

³*Facultad de Ciencias Físicas y Matemáticas, Departamento de Estadística, Universidad de Concepción, Concepción, Chile*

We define the bias term for the estimator for the gradient at the k -th iteration of CSPSA as

$$\mathbf{b}_k(\hat{\boldsymbol{\mu}}_k) = \mathbb{E}(\hat{\mathbf{g}}_k(\hat{\boldsymbol{\mu}}_k) - \mathbf{g}(\hat{\boldsymbol{\mu}}_k) | \hat{\boldsymbol{\mu}}_k), \quad (1)$$

and the error term for the k -th iteration of CSPSA as

$$\mathbf{e}_k(\hat{\boldsymbol{\mu}}_k) = \hat{\mathbf{g}}_k(\hat{\boldsymbol{\mu}}_k) - \mathbb{E}(\hat{\mathbf{g}}_k(\hat{\boldsymbol{\mu}}_k) | \hat{\boldsymbol{\mu}}_k). \quad (2)$$

We also define the set $\mathcal{F}_k = \sigma(\{\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \dots, \hat{\boldsymbol{\mu}}_k\})$. Because the Δ_k are independent of $\mathcal{F}_k$, we have that

$$\mathbb{E}(\hat{\mathbf{g}}_k(\hat{\boldsymbol{\mu}}_k) | \mathcal{F}_k) = \mathbb{E}(\hat{\mathbf{g}}_k(\hat{\boldsymbol{\mu}}_k) | \hat{\boldsymbol{\mu}}_k). \quad (3)$$

Let us now consider the real vector $\mathbf{x} = (\text{Re}(z), \text{Im}(z))$. There is an isomorphism between the space of vectors $\mathbf{x}$ and the space of vectors $\boldsymbol{\mu} = (z, z^*)$. Indeed, the matrix

$$\mathbf{J} = \begin{bmatrix} \mathbb{I}_n & i\mathbb{I}_n \\ \mathbb{I}_n & -i\mathbb{I}_n \end{bmatrix} \quad (4)$$

satisfies $\boldsymbol{\mu} = \mathbf{J}\mathbf{x}$ and is invertible. Therefore, we can regard any function of $\boldsymbol{\mu}$ as a function of $\mathbf{x}$ [1].

Lemma 1. *Let W be a vector space on $\mathbb{R}$, and $U \subseteq \mathfrak{C}^n$ an open subset. Let $f : U \rightarrow W$ such that $f \in \mathcal{C}^K$, $K \geq 1$. Let us choose $\boldsymbol{\mu}_0 \in U$ and $r > 0$ such that $B_r(\boldsymbol{\mu}_0) \subset U$. Let us choose $\mathbf{h} \in \mathfrak{C}^n$ such that $\|\mathbf{h}\| < r$. For positive integers $j < K$, we define the differential operator $\mathcal{D}_{\mathbf{h}}^{(j)}$ over f as*

$$(\mathcal{D}_{\mathbf{h}}^{(j)} f)(\boldsymbol{\mu}_0) = \sum_{i_1, \dots, i_j} h_{i_1} \dots h_{i_j} \frac{\partial^j f}{\partial \mu_{i_1} \dots \partial \mu_{i_j}}(\boldsymbol{\mu}_0). \quad (5)$$

Then,

$$f(\boldsymbol{\mu}_0 + \mathbf{h}) = \sum_{j=0}^2 \frac{1}{j!} (\mathcal{D}_{\mathbf{h}}^{(j)} f)(\boldsymbol{\mu}_0) + R_{2,\mathbf{h}}(\boldsymbol{\mu}_0), \quad (6)$$

where

$$R_{2,\mathbf{h}}(\boldsymbol{\mu}_0) = \int_0^1 (1-t) [(\mathcal{D}_{\mathbf{h}}^{(2)} f)(\boldsymbol{\mu}_0 + t\mathbf{h}) - (\mathcal{D}_{\mathbf{h}}^{(2)} f)(\boldsymbol{\mu}_0)] dt, \quad (7)$$

Proof. Because we can regard a function of $\boldsymbol{\mu}$ as a function of $\mathbf{x}$, we have $f(\boldsymbol{\mu}_0 + \mathbf{h}) = f(\mathbf{x}_0 + \mathbf{k})$. From theorem 5.8 in [2], we have that

$$f(\mathbf{x}_0 + \mathbf{k}) = \sum_{j=0}^2 \frac{1}{j!} (D_{\mathbf{k}}^{(j)} f)(\mathbf{x}_0) + P_{2,\mathbf{k}}(\mathbf{x}_0) \quad (8)$$

with

$$P_{2,\mathbf{k}}(\mathbf{x}_0) = \int_0^1 (1-t) [(D_{\mathbf{k}}^{(2)} f)(\mathbf{x}_0 + t\mathbf{k}) - (D_{\mathbf{k}}^{(2)} f)(\mathbf{x}_0)] dt \quad (9)$$

where $\mathbf{x}_0 = J^{-1}\boldsymbol{\mu}_0$, $\mathbf{k} = J^{-1}\mathbf{h}$ and the real differential operator $D_{\mathbf{k}}^{(j)}$ onto function f is

$$(D_{\mathbf{k}}^{(j)} f)(\mathbf{x}) = \sum_{i_1, \dots, i_j}^{2n} k_{i_1} \dots k_{i_j} (\partial_{x_{i_1}} \dots \partial_{x_{i_j}} f)(\mathbf{x}). \quad (10)$$

Then,

$$\begin{aligned} f(\mathbf{x}_0 + \mathbf{k}) &= f(\mathbf{x}_0) + \sum_{i=1}^n k_i \frac{\partial f}{\partial x_i}(\mathbf{x}_0) + \frac{1}{2} \sum_{i=1}^n k_i^2 \frac{\partial^2 f}{\partial x_i^2}(\mathbf{x}_0) \\ &\quad + \sum_{i < j} k_i k_j \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0) + P_{2,\mathbf{k}}(\mathbf{x}_0) \\ &= f(\mathbf{x}_0) + \mathbf{k}^T \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}_0) \\ &\quad + \frac{1}{2} \mathbf{k}^T \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T}(\mathbf{x}_0) \mathbf{k} + P_{2,\mathbf{k}}(\mathbf{x}_0). \end{aligned}$$

Because we have $f(\mathbf{x}_0) = f(\boldsymbol{\mu}_0)$:

$$\begin{aligned} f(\mathbf{x}_0 + \mathbf{k}) &= f(\boldsymbol{\mu}_0) + \mathbf{h}^T (J^{-1})^T J^T \frac{\partial f}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}_0) \\ &\quad + \frac{1}{2} \mathbf{h}^T (J^{-1})^T J^T \frac{\partial^2 f}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^T}(\boldsymbol{\mu}_0) J J^{-1} \mathbf{h} + P_{2,\mathbf{k}}(\mathbf{x}_0) \\ &= f(\boldsymbol{\mu}_0) + \mathbf{h}^T \frac{\partial f}{\partial \boldsymbol{\mu}}(\boldsymbol{\mu}_0) \\ &\quad + \frac{1}{2} \mathbf{h}^T \frac{\partial^2 f}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^T}(\boldsymbol{\mu}_0) \mathbf{h} + P_{2,\mathbf{k}}(\mathbf{x}_0) \\ &= \sum_{j=0}^2 \frac{1}{j!} (D_{\mathbf{h}}^{(j)} f)(\boldsymbol{\mu}_0) + P_{2,\mathbf{k}}(\mathbf{x}_0). \end{aligned}$$

Also,

$$\begin{aligned}
P_{2,\mathbf{k}}(\mathbf{x}_0) &= \int_0^1 (1-t)((D_{\mathbf{k}}^2 f)(\mathbf{x}_0 + t\mathbf{k}) - (D_{\mathbf{k}}^2 f)(\mathbf{x}_0))dt \\
&= 2 \int_0^1 (1-t) \left[\frac{1}{2} \sum_{i=0}^2 k_i^2 \left(\frac{\partial^2 f}{\partial x_i^2}(\mathbf{x}_0 + t\mathbf{k}) - \frac{\partial^2 f}{\partial x_i^2}(\mathbf{x}_0) \right) + \sum_{i < j} h_{[R]j} h_{[R]i} \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0 + t\mathbf{k}) - \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0) \right) \right] dt \\
&= \int_0^1 (1-t) \mathbf{k}^T \left(\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T}(\mathbf{x}_0 + t\mathbf{k}) - \frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T}(\mathbf{x}_0) \right) \mathbf{k} dt \\
&= \int_0^1 (1-t) \mathbf{h}^T \left(\frac{\partial^2 f}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^T}(\boldsymbol{\mu}_0 + t\mathbf{h}) - \frac{\partial^2 f}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^T}(\boldsymbol{\mu}_0) \right) \mathbf{h} dt \\
&= \int_0^1 (1-t)((D_{\mathbf{h}}^2 f)(\boldsymbol{\mu}_0 + t\mathbf{h}) - (D_{\mathbf{h}}^2 f)(\boldsymbol{\mu}_0))dt \\
&= R_{2,\mathbf{h}}(\boldsymbol{\mu}_0).
\end{aligned}$$

Therefore,

$$f(\boldsymbol{\mu}_0 + \mathbf{h}) = \sum_{j=0}^2 \frac{1}{j!} (D_{\mathbf{h}}^{(j)} f)(\boldsymbol{\mu}_0) + R_{2,\mathbf{h}}(\boldsymbol{\mu}_0). \quad (11)$$

□

Theorem 1. For α_0, α_1 and α_2 positive real constants and $\Omega = \{\omega\}$ the sample space that generates the sequence $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, \dots$, consider all $k \geq K$ for some $K < \infty$, and suppose that for each such k the $\Delta_{k,i}$ are independent and identically distributed and such that $\mathbb{E}(\Delta_{k,i}) = 0$, with $|\Delta_{k,i}| \leq \alpha_0$ a.s., $\mathbb{E}(|\Delta_{k,i}^{-1}|) \leq \alpha_1$ and $\mathbb{E}[e^{2i\phi_{k,i}}] = 0$, where $\phi_{k,i}$ is the phase of

$\Delta_{k,i}$ in its polar decomposition. For example, $\Delta_{k,i}$ can be taken as symmetrically distributed about 0. For almost all $\hat{\boldsymbol{\mu}}_k$ (at each $k \geq K$) suppose that for all $\boldsymbol{\mu}$ in an open neighborhood of $\hat{\boldsymbol{\mu}}_k$, that is not a function of k or ω , $f \in \mathcal{C}^2(\mathbb{C}^n)$ with $|\partial_{\mu_i} \partial_{\mu_j} f(\boldsymbol{\mu})| \leq \alpha_2$. Then for almost all $\omega \in \Omega$

$$\mathbf{b}_k(\hat{\boldsymbol{\mu}}_k) = \mathcal{O}(c_k). \quad (12)$$

Proof. We apply lemma 1 to $f(\boldsymbol{\mu})$. Besides, because $|\partial_i \partial_j f(\boldsymbol{\mu})|$ is bounded, by hypothesis, $\mathbb{E}(|\partial_i \partial_j f(\boldsymbol{\mu})|)$ is well defined. For all $l \in \{1, \dots, p\}$:

$$\mathbf{b}_{k,l}(\hat{\boldsymbol{\mu}}_k) = \mathbb{E}(\hat{g}_{k,l}(\hat{\boldsymbol{\mu}}_k) - g_l(\hat{\boldsymbol{\mu}}_k) | \hat{\boldsymbol{\mu}}_k) = \mathbb{E} \left(\left[\frac{f(\hat{\mathbf{z}}_k + c_k \boldsymbol{\Lambda}_k) - f(\hat{\mathbf{z}}_k - c_k \boldsymbol{\Lambda}_k)}{2c_k \Delta_{k,l}^*} - \frac{\partial f}{\partial z_l^*}(\hat{\boldsymbol{\mu}}_k) \right] \middle| \hat{\boldsymbol{\mu}}_k \right) \quad (13)$$

where $\boldsymbol{\Lambda} = (\Delta, \Delta^*)$.

When we expand $f(\hat{\boldsymbol{\mu}}_k \pm c_k \boldsymbol{\Lambda}_k)$, the terms with even powers of $\boldsymbol{\Lambda}_k$ will cancel each other.

$$\begin{aligned}
\mathbf{b}_{k,l}(\hat{\boldsymbol{\mu}}_k) &= \mathbb{E} \left(\left[2 \sum_{i=1}^{2n} \frac{c_k \Lambda_{k,i}}{2c_k \Delta_{k,l}^*} \frac{\partial f}{\partial \mu_i}(\hat{\boldsymbol{\mu}}_k) \right. \right. \\
&\quad \left. \left. + \frac{R_{2,c_k \boldsymbol{\Lambda}_k}(\hat{\boldsymbol{\mu}}_k) - R_{2,-c_k \boldsymbol{\Lambda}_k}(\hat{\boldsymbol{\mu}}_k)}{2c_k \Delta_{k,l}^*} - \frac{\partial f}{\partial z_l^*}(\hat{\boldsymbol{\mu}}_k) \right] \middle| \hat{\boldsymbol{\mu}}_k \right) \\
&= \mathbb{E} \left(\left[\sum_{i=1}^n \frac{\Delta_{k,i}}{\Delta_{k,l}^*} \frac{\partial f}{\partial z_i}(\hat{\boldsymbol{\mu}}_k) + \sum_{i \neq l}^n \frac{\Delta_{k,i}^*}{\Delta_{k,l}^*} \frac{\partial f}{\partial z_i^*}(\hat{\boldsymbol{\mu}}_k) \right. \right. \\
&\quad \left. \left. + \frac{R_{2,c_k \boldsymbol{\Lambda}_k}(\hat{\boldsymbol{\mu}}_k) - R_{2,-c_k \boldsymbol{\Lambda}_k}(\hat{\boldsymbol{\mu}}_k)}{2c_k \Delta_{k,l}^*} \right] \middle| \hat{\boldsymbol{\mu}}_k \right).
\end{aligned}$$

Because $\partial_i f(\hat{\boldsymbol{\mu}}_k)$ is completely determined by $\hat{\boldsymbol{\mu}}_k$,

$$\mathbb{E} \left(\sum_{i \neq l}^n \frac{\Delta_{k,i}^*}{\Delta_{k,l}^*} \frac{\partial f}{\partial z_i^*}(\hat{\boldsymbol{\mu}}_k) \middle| \hat{\boldsymbol{\mu}}_k \right) = \mathbb{E} \left(\sum_{i \neq l}^n \frac{\Delta_{k,i}^*}{\Delta_{k,l}^*} \middle| \hat{\boldsymbol{\mu}}_k \right) \frac{\partial f}{\partial z_i^*}(\hat{\boldsymbol{\mu}}_k). \quad (14)$$

Besides, the $\{\Delta_{k,i}\}$ are independently distributed, and are independent of $\hat{\boldsymbol{\mu}}_k$, so $\mathbb{E} \left(\sum_{i \neq l}^n \frac{\Delta_{k,i}^*}{\Delta_{k,l}^*} \middle| \hat{\boldsymbol{\mu}}_k \right) = \mathbb{E}((\Delta_{k,l}^*)^{-1}) \sum_{i \neq l}^n \mathbb{E}(\Delta_{k,i}^*)$. Then

$$|b_{k,l}(\hat{\mu}_k)| \leq \left| \mathbb{E} \left(\frac{\Delta_{k,l}}{\Delta_{k,l}^*} \right) \frac{\partial f}{\partial z_l^*}(\hat{\mu}_k) \right| + \left| \mathbb{E}((\Delta_{k,l}^*)^{-1}) \sum_{i \neq l}^n \left(\mathbb{E}(\Delta_{k,i}) \frac{\partial f}{\partial z_i}(\hat{\mu}_k) + \mathbb{E}(\Delta_{k,i}^*) \frac{\partial f}{\partial z_i^*}(\hat{\mu}_k) \right) \right| \\ + \left| \mathbb{E} \left(\frac{R_{2,c_k \Lambda_k}(\hat{\mu}_k) - R_{2,-c_k \Lambda_k}(\hat{\mu}_k)}{2c_k \Delta_{k,l}^*} \right) \hat{\mu}_k \right|.$$

Because $\mathbb{E}(\Delta_{ik}) = \mathbb{E}(\Delta_{ik}^*) = 0$,

$$|b_{k,l}(\hat{\mu}_k)| \leq \left| \mathbb{E}(e^{2i\phi_{kl}}) \frac{\partial f}{\partial z_l^*}(\hat{\mu}_k) \right| \\ + \mathbb{E} \left(\left| \frac{R_{2,c_k \Lambda_k}(\hat{\mu}_k) - R_{2,-c_k \Lambda_k}(\hat{\mu}_k)}{2c_k \Delta_{k,l}^*} \right| \hat{\mu}_k \right).$$

Therefore $\mathbb{E}(e^{2i\phi_{kl}}) = 0$,

$$|b_{k,l}(\hat{\mu}_k)| \leq \mathbb{E} \left(\left| \int_0^1 \frac{1-t}{2c_k \Delta_{k,l}} \left[(\mathcal{D}_{c_k \Lambda_k}^2 f)(\hat{\mu}_k + tc_k \Lambda_k) \right. \right. \right. \\ \left. \left. \left. - (\mathcal{D}_{c_k \Lambda_k}^2 f)(\hat{\mu}_k) - (\mathcal{D}_{-c_k \Lambda_k}^2 f)(\hat{\mu}_k - tc_k \Lambda_k) \right. \right. \right. \\ \left. \left. \left. + (\mathcal{D}_{-c_k \Lambda_k}^2 f)(\hat{\mu}_k) \right] dt \right| \hat{\mu}_k \right).$$

When we expand the differential operators, we can see that $(\mathcal{D}_{c_k \Lambda_k}^2 f)(\hat{\mu}_k) = (\mathcal{D}_{-c_k \Lambda_k}^2 f)(\hat{\mu}_k)$ and that $(\mathcal{D}_{-c_k \Lambda_k}^2 f)(\hat{\mu}_k - tc_k \Lambda_k) = (\mathcal{D}_{c_k \Lambda_k}^2 f)(\hat{\mu}_k - tc_k \Lambda_k)$. Therefore,

$$|b_{k,l}(\hat{\mu}_k)| \leq \mathbb{E} \left(\left| \int_0^1 \frac{1-t}{2c_k \Delta_{k,l}} \left[(\mathcal{D}_{c_k \Lambda_k}^2 f)(\hat{\mu}_k + tc_k \Lambda_k) \right. \right. \right. \\ \left. \left. \left. + (\mathcal{D}_{c_k \Lambda_k}^2 f)(\hat{\mu}_k - tc_k \Lambda_k) \right] dt \right| \hat{\mu}_k \right) \\ \leq \mathbb{E} \left(\left| \int_0^1 \frac{1-t}{2c_k |\Delta_{k,l}|} \left[|(\mathcal{D}_{c_k \Lambda_k}^2 f)(\hat{\mu}_k + tc_k \Lambda_k)| \right. \right. \right. \\ \left. \left. \left. + |(\mathcal{D}_{c_k \Lambda_k}^2 f)(\hat{\mu}_k - tc_k \Lambda_k)| \right] dt \right| \hat{\mu}_k \right).$$

Each differential operator consists of a sum over $4n^2$ terms, each the product of two components of $c_k \Lambda_k$ and a second order derivative. From hypothesis, we have that those quantities are bounded by $|\Delta_{k,i}| \leq \alpha_0$ y $|\partial_i \partial_j f(z)| \leq \alpha_2$, so

$$|b_{k,l}(\hat{\mu}_k)| \leq \int_0^1 \frac{1-t}{2c_k} \mathbb{E}|\Delta_{k,l}^{-1}| [8n^2 c_k^2 \alpha_0^2 \alpha_2] dt. \quad (15)$$

Considering that $\mathbb{E}|\Delta_{k,i}^{-1}| \leq \alpha_1$,

$$|b_{k,l}(\hat{\mu}_k)| \leq 2c_k n^2 \alpha_0^2 \alpha_1 \alpha_2. \quad (16)$$

□

Theorem 2. Let the conditions of theorem 1 and the following assumptions hold:

- (i): $a_k, c_k > 0 \forall k$; $\lim_{k \rightarrow \infty} (|a_k| + |c_k|) = 0$; $\sum_{k=1}^{\infty} a_k = \infty$; $\sum_{k=1}^{\infty} (a_k/c_k)^2 < \infty$,
- (ii): For some $\beta_0, \beta_1, \beta_2 > 0$, $\forall k$, $\mathbb{E}(\varepsilon_{k,\pm}^2) \leq \beta_0$, $\mathbb{E}([f(\hat{\mu}_k \pm c_k \Lambda_k)]^2) \leq \beta_1$, and $\mathbb{E}(\Delta_{k,i}^{-2}) \leq \beta_2$.

(iii): $\|\hat{\mu}_k\| < \infty$ for almost all k ,

(iv): $z(t) = \tilde{z}$ is an asymptotically stable solution of the differential equation $dz(t)/dt = -g(\mu)$,

(v): Consider the domain of attraction $D(\tilde{z}) = \{z_0 : \lim_{t \rightarrow \infty} z(t|z_0) = \tilde{z}\}$ where $z(t|z_0)$ denotes the solutions to the differential equation of A4 based on initial conditions $z(0) = z_0$. There exists a closed set $S \subseteq D(\tilde{z})$ such that $\hat{z}_k \in S$ infinitely often for almost all sample points,

Then

$$\lim_{k \rightarrow \infty} \hat{z}_k = \tilde{z} \quad (17)$$

for almost all $\omega \in \Omega$.

Proof. According to [3], SPSA converges to the solution of the optimisation problem if

1. $\|b_k(\hat{x}_k)\| < \infty$ and $\lim_{k \rightarrow \infty} b_k(\hat{x}_k) = 0$ almost surely
2. $\lim_{k \rightarrow \infty} P(\sup_{m \geq k} \|\sum_{i=k}^m a_i e_i(\hat{x}_i)\| \geq \eta) = 0$ for all $\eta > 0$

Here, $\hat{x}_k$ is the estimate at the k th iteration of the vector $\tilde{x}$ which minimises the function $f(x)$, just like $\hat{\mu}_k$ and $\tilde{\mu}$ for $f(\mu)$, only for real numbers. The functions $b_k(\hat{x}_k)$ and $e_k(\hat{x}_k)$ are defined in a similar way, and assumptions similar to those in theorem 2 are made.

We will show that similar statements for complex numbers imply the above, and thus the convergence of SPSA, and that in turn implies the convergence of CSPA.

Firstly, we will prove that the following hold:

1. $\|b_k(\hat{\mu}_k)\| < \infty$ and $\lim_{k \rightarrow \infty} b_k(\hat{\mu}_k) = 0$ almost surely
2. $\lim_{k \rightarrow \infty} P(\sup_{m \geq k} \|\sum_{i=k}^m a_i e_i(\hat{\mu}_i)\| \geq \eta) = 0$ for all $\eta > 0$

It is shown in theorem 1 that $|b_k(\hat{\mu}_{k,l})|$ has an upper bound directly proportional to c_k . Due to condition (i), $\lim_{k \rightarrow \infty} c_k = 0$. This proves condition 1.

To prove condition 2, we notice that the sequence $\{\sum_{i=k}^m a_i e_i(\hat{\mu}_i)\}_{m \geq k}$ is a martingale. Indeed,

$$\mathbb{E} \left(\sum_{i=k}^{m+1} a_i e_i(\hat{\mu}_i) - \sum_{i=k}^m a_i e_i(\hat{\mu}_i) \middle| \hat{\mu}_m \right) = \mathbb{E}(a_{m+1} e_{m+1}(\hat{\mu}_{m+1}) | \hat{\mu}_m) = a_{m+1} (\mathbb{E}(\hat{g}_{m+1}(\hat{\mu}_{m+1}) | \hat{\mu}_m) - \mathbb{E}(\mathbb{E}(\hat{g}_{m+1}(\hat{\mu}_{m+1}) | \hat{\mu}_{m+1}) | \hat{\mu}_m)).$$

But $\mathbb{E}(\mathbb{E}(\hat{g}_{m+1}(\hat{\mu}_{m+1}) | \hat{\mu}_{m+1}) | \hat{\mu}_m) = \mathbb{E}(\hat{g}_{m+1}(\hat{\mu}_{m+1}) | \hat{\mu}_{m+1})$ because $\mathcal{F}_m \subset \mathcal{F}_{m+1}$, so

$$\mathbb{E} \left(\sum_{i=k}^{m+1} a_i e_i(\hat{\mu}_i) - \sum_{i=k}^m a_i e_i(\hat{\mu}_i) \middle| \hat{\mu}_m \right) = 0. \quad (18)$$

Therefore, $\{\|\sum_{i=k}^m a_i e_i(\hat{\mu}_i)\|\}_{m \geq k}$ is a submartingale. By Doob's inequality, if we have a submartingale $\{X_n\}$, for all $\eta > 0$ and $p \geq 1$ we have

$$P \left(\max_{1 \leq i \leq n} X_i \geq \eta \right) \leq \eta^{-p} \mathbb{E}(|X_n|^p). \quad (19)$$

Then,

$$P \left(\sup_{m \geq k} \left\| \sum_{i=k}^m a_i e_i(\hat{\mu}_i) \right\| \geq \eta \right) \leq \eta^{-2} \mathbb{E} \left(\left\| \sum_{i=k}^{\infty} a_i e_i(\hat{\mu}_i) \right\|^2 \right). \quad (20)$$

Because $\mathbb{E}(e_i^T e_j) = \mathbb{E}(e_i^T \mathbb{E}(e_j | \hat{z}_i)) = 0$ when $i < j$ we have

$$P \left(\sup_{m \geq k} \left\| \sum_{i=k}^m a_i e_i(\hat{\mu}_i) \right\| \geq \eta \right) \leq \eta^{-2} \sum_{i=k}^{\infty} a_i \mathbb{E}(\|e_i(\hat{\mu}_i)\|^2). \quad (21)$$

We can see that

$$\begin{aligned} \mathbb{E}(|\hat{g}_{k,l}(\hat{\mu}_k)|^2) &= \frac{1}{4} \mathbb{E}[\mathbb{E}(|(c_k \Delta_{k,l}^*)^{-1} (f(\mu_k + c_k \Lambda_k) \\ &\quad - f(\mu_k - c_k \Lambda_k) + \varepsilon_{k,+} - \varepsilon_{k,-})|^2 | \mu_k, \Lambda_k)] \\ &= \frac{1}{4} \mathbb{E}[|c_k \Delta_{k,l}^*|^{-2} \mathbb{E}(|f(\mu_k + c_k \Lambda_k) \\ &\quad - f(\mu_k - c_k \Lambda_k) + \varepsilon_{k,+} - \varepsilon_{k,-}|^2 | \mu_k, \Lambda_k)] \\ &\leq \frac{1}{2} \mathbb{E}[|c_k \Delta_{k,l}^*|^{-2} \mathbb{E}(|f(\mu_k + c_k \Lambda_k) \\ &\quad - f(\mu_k - c_k \Lambda_k)|^2 + |\varepsilon_{k,+} - \varepsilon_{k,-}|^2 | \mu_k, \Lambda_k)]. \end{aligned}$$

From condition (ii), $\mathbb{E}(\varepsilon_{k\pm})^2 \leq \beta_0$ and $\mathbb{E}f(\mu_k \pm c_k \Lambda_k)^2 \leq \beta_1$, so

$$\mathbb{E}(|\hat{g}_{k,l}(\hat{\mu}_k)|^2) \leq 2\mathbb{E}(|c_k \Delta_{k,l}|^{-2} (\beta_0 + \beta_1)). \quad (22)$$

We also have that $\mathbb{E}(|\Delta_{k,l}|^{-2}) \leq \beta_2$, thus

$$\mathbb{E}(|\hat{g}_{k,l}(\hat{\mu}_k)|^2) \leq 2(\beta_0 + \beta_1)\beta_2 c_k^{-2}. \quad (23)$$

On the other hand, we have that

$$\begin{aligned} \mathbb{E}(\|e_k(\hat{\mu}_k)\|^2) &= \mathbb{E}(\|\hat{g}_k(\hat{\mu}_k) - \mathbb{E}(\hat{g}_k(\hat{\mu}_k) | \hat{\mu}_k)\|^2) \\ &\leq \mathbb{E}(\|\hat{g}_k(\hat{\mu}_k)\| + \|\mathbb{E}(\hat{g}_k(\hat{\mu}_k) | \hat{\mu}_k)\|)^2 \\ &\leq 2\mathbb{E}(\|\hat{g}_k(\hat{\mu}_k)\|^2) + 2\mathbb{E}(\|\mathbb{E}(\hat{g}_k(\hat{\mu}_k) | \hat{\mu}_k)\|^2) \\ &= 4 \sum_{l=1}^n \mathbb{E}(|\hat{g}_{k,l}(\hat{\mu}_k)|^2) \\ &\leq 8n(\beta_0 + \beta_1)\beta_2 c_k^{-2}. \end{aligned}$$

Substituting in (21), we obtain

$$P \left(\sup_{m \geq k} \left\| \sum_{i=k}^m a_i e_i(\hat{\mu}_i) \right\| \geq \eta \right) \leq \frac{8n}{\eta^2} (\beta_0 + \beta_1) \beta_2 \sum_{i=k}^{\infty} \left(\frac{a_i}{c_i} \right)^2 \quad (24)$$

Condition (i) says that $\sum_{i=0}^{\infty} \left(\frac{a_i}{c_i} \right)^2 < \infty$, so

$\lim_{k \rightarrow \infty} \sum_{i=k}^{\infty} \left(\frac{a_i}{c_i} \right)^2 = 0$. Statement 2 follows.

We will show that statements 1 and 2 for CSPSA satisfy the conditions for Theorem 2.3.1 in [4]. Because we can regard any function of μ as a function of x [1], we have that $b_k(\hat{\mu}_k) = b_k(\hat{x}_k)$ and $e_k(\hat{\mu}_k) = e_k(\hat{x}_k)$. According to [4], this implies that the algorithm

$$\hat{x}_{k+1} = \hat{x}_k - a_k [\nabla f(\hat{x}_k) + b_k(\hat{x}_k) + e_k(\hat{x}_k)] \quad (25)$$

satisfies

$$\lim_{k \rightarrow \infty} \hat{x}_k = \tilde{x} \quad (26)$$

with $\tilde{x} = (\text{Re}(\tilde{z}), \text{Im}(\tilde{z}))$. Using (2), we can rewrite the CSPSA algorithm as

$$\hat{z}_{k+1} = \hat{z}_k - a_k [\partial_{z^*} f(\hat{\mu}_k) + b_k(\hat{\mu}_k) + e_k(\hat{\mu}_k)]. \quad (27)$$

Because $\partial_{z^*} = \partial_x + i\partial_y$, we can identify algorithm (25) with (27), so $\hat{x}_k = (\text{Re}(\hat{z}_k), \text{Im}(\hat{z}_k))$. Therefore,

$$\lim_{k \rightarrow \infty} \hat{z}_k = \tilde{z}. \quad (28)$$

□

- [3] J. C. Spall, [IEEE Trans. Autom. Control](#) **37**, 332 (1992).
- [4] H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems* (Springer-Verlag New York, 1978).