Monte-Carlo SURE: A Black-Box Optimization of Regularization Parameters for General Denoising Algorithms - Supplementary Material

Sathish Ramani*, *Student Member*, Thierry Blu, *Senior Member*, and Michael Unser, *Fellow*

This material supplements some sections of the paper entitled "Monte-Carlo SURE: A Black-Box Optimization of Regularization Parameters for General Denoising Algorithms". Here, we elaborate on the solution to the differentiability issue associated with the Monte-Carlo divergence estimation proposed (in Theorem 2) in the paper. Firstly, we verify the validity of the Taylor expansion-based argumentation of Theorem 2 for algorithms like total-variation denoising (TVD). Following that, we give a proof of the second part of Theorem 2 which deals with a weaker hypothesis (using tempered distributions) of the problem.

I. Total-Variation Denoising: Verification of Taylor Expansion-based hypothesis

In the paper, we considered the discrete-domain formulation for total-variation denoising (TVD) where we minimize the cost

$$\mathcal{J}_{\mathrm{TV}}(\mathbf{u}) = \|\mathbf{y}-\mathbf{u}\|^2 + \lambda\,\mathrm{TV}\{\mathbf{u}\}, \tag{1}$$

where $\mathrm{TV}\{\mathbf{u}\} = \sum_k \sqrt{(\mathbf{D}_h\mathbf{u})[k]^2 + (\mathbf{D}_v\mathbf{u})[k]^2}$ is the discrete 2D total variation norm and $\mathbf{D}_h$ and $\mathbf{D}_v$ are matrices corresponding to the first order finite difference in the horizontal and vertical directions, respectively. We will concentrate only on the bounded-optimization (half-quadratic) algorithm developed in [1]. However, since a typical implementation of TVD will always involve finite differences in place of continuous domain derivatives the analysis can be easily extended to other algorithms including the variant in [1] and those based on Euler-Lagrange equations. We show that the bounded-optimization algorithm for TVD admits first and second order derivatives with respect to the data $\mathbf{y}$ and therefore satisfies the stronger hypothesis (Taylor expansion-based) of Theorem 2.

The TVD algorithm is described by the following recursive equation [1]: the signal estimate at iteration $k+1$ denoted by the $N\times 1$ vector $\mathbf{f}_{\boldsymbol{\lambda}}^{k+1}$ is obtained by solving the linear system

$$\mathbf{M}^k\mathbf{f}_{\boldsymbol{\lambda}}^{k+1} = \mathbf{y}, \tag{2}$$

where $\mathbf{M}^k$ is the $N\times N$ system matrix at iteration k given by

$$\mathbf{M}^k = \mathbf{I} + \mathbf{D}_h^{\mathrm{T}}\Lambda^k\mathbf{D}_h + \mathbf{D}_v^{\mathrm{T}}\Lambda^k\mathbf{D}_v, \tag{3}$$

and $\Lambda^k = \mathrm{diag}\{w_i^k;\ i=1,2,\ldots,N\}$, with

$$w_i^k = \frac{\lambda}{2\sqrt{(\mathbf{D}_h\mathbf{f}_{\boldsymbol{\lambda}}^k)_i^2 + (\mathbf{D}_v\mathbf{f}_{\boldsymbol{\lambda}}^k)_i^2 + \kappa}}, \tag{4}$$

where $(\mathbf{D}_*\mathbf{f}_{\boldsymbol{\lambda}}^k)_i$ is the i^{th} element of the vector $\mathbf{D}_*\mathbf{f}_{\boldsymbol{\lambda}}^k$ and $\kappa > 0$ is a small constant that prevents the denominator of w_i^k going to zero (or else the algorithm $\mathbf{f}_{\boldsymbol{\lambda}}$ itself may become numerically unstable).

Email: sathish.ramani@epfl.ch, thierry.blu@m4x.org, michael.unser@epfl.ch.

Differentiating (2) with respect to $\mathbf{y}$, we obtain

$$\frac{\partial \mathbf{M}^k}{\partial \mathbf{y}} \mathbf{f}_{\boldsymbol{\lambda}}^{k+1} + \mathbf{M}^k \underbrace{\frac{\partial \mathbf{f}_{\boldsymbol{\lambda}}^{k+1}}{\partial \mathbf{y}}}_{\mathbf{J}_{\mathbf{f}_{\boldsymbol{\lambda}}}^{k+1}} = \mathbf{I}, \tag{5}$$

where $\mathbf{J}_{\mathbf{f}_{\boldsymbol{\lambda}}}^{k+1}$ is the Jacobian matrix of $\mathbf{f}_{\boldsymbol{\lambda}}^{k+1}$ at iteration $k+1$. If M_{mn}^k represents the mn^{th} element of $\mathbf{M}^k$ and y_m and f_m^k represent the m^{th} elements of $\mathbf{y}$ and $\mathbf{f}_{\boldsymbol{\lambda}}^k$, respectively, then using Einstein's summation notation (repeated indices will be summed over unless they appear on both sides of an equation) (5) can be written as

$$\frac{\partial M_{mn}^k}{\partial y_p} \mathrm{f}_n^{k+1} + M_{mn}^k \frac{\partial \mathrm{f}_n^{k+1}}{\partial y_p} = \delta_{mp}, \tag{6}$$

where, for example, the index n is summer over in both the terms on the L.H.S of the above equation.

Differentiating (6) a second time, we obtain

$$\frac{\partial^2 M_{mn}^k}{\partial y_l \partial y_p} \mathrm{f}_n^{k+1} + \frac{\partial M_{mn}^k}{\partial y_p} \frac{\partial \mathrm{f}_n^{k+1}}{\partial y_l} + \frac{\partial M_{mn}^k}{\partial y_l} \frac{\partial \mathrm{f}_n^{k+1}}{\partial y_p} + M_{mn}^k \frac{\partial \mathrm{f}_n^{k+1}}{\partial y_l \partial y_p} = 0. \tag{7}$$

It is clear that the $N \times N \times N$ tensor $\mathbf{r} = \left\{ \frac{\partial \mathrm{f}_n^{k+1}}{\partial y_l \partial y_p} \right\}_{n,l,p=1}^{N}$ is the desired second derivative in the Taylor expansion in Theorem 2. We will show that for a given (l,p), all the terms in (7) are well-defined, so that the $N \times 1$ vector $\mathbf{r}_{lp} = \frac{\partial \mathbf{f}_{\boldsymbol{\lambda}}{}^{k+1}}{\partial y_l \partial y_p}$ can be obtained by solving a linear system.

Firstly, we analyze $\frac{\partial M_{mn}^k}{\partial y_p}$ which is given by

$$\frac{\partial M_{mn}^k}{\partial y_p} = \left(D_{h_{qm}} D_{h_{qn}} + D_{v_{qm}} D_{v_{qn}} \right) \frac{\partial w_q^k}{\partial y_p}, \tag{8}$$

with

$$\frac{\partial w_q^k}{\partial y_p} = -\left(\frac{2}{\lambda}\right)^2 (w_q^k)^3 \left((\mathbf{D}_h \mathbf{f}_{\boldsymbol{\lambda}}^k)_q D_{h_{qi}} + (\mathbf{D}_v \mathbf{f}_{\boldsymbol{\lambda}}^k)_q D_{v_{qi}} \right) \frac{\partial \mathrm{f}_i^k}{\partial y_p}, \tag{9}$$

where the index q is not summed over on the R.H.S of the above equation.

Similarly,

$$\frac{\partial^2 M_{mn}^k}{\partial y_l \partial y_p} = \left(D_{h_{qm}} D_{h_{qn}} + D_{v_{qm}} D_{v_{qn}} \right) \frac{\partial^2 w_q^k}{\partial y_l \partial y_p}, \tag{10}$$

where

$$\frac{\partial^2 w_q^k}{\partial y_l \partial y_p} = -\left(\frac{2}{\lambda}\right)^2 (w_q^k)^3 \begin{bmatrix} \frac{1}{3}\left(\frac{\lambda}{2}\right)^2 (w_q^k)^{-4} \frac{\partial w_q^k}{\partial y_l} \frac{\partial w_q^k}{\partial y_p} \\ -\left((\mathbf{D}_h \mathbf{f}_{\boldsymbol{\lambda}}^k)_q D_{h_{qi}} D_{h_{qj}} + (\mathbf{D}_v \mathbf{f}_{\boldsymbol{\lambda}}^k)_q D_{v_{qi}} D_{v_{qj}} \right) \frac{\partial \mathrm{f}_i^k}{\partial y_l} \frac{\partial \mathrm{f}_j^k}{\partial y_p} \\ -\left((\mathbf{D}_h \mathbf{f}_{\boldsymbol{\lambda}}^k)_q D_{h_{qi}} + (\mathbf{D}_v \mathbf{f}_{\boldsymbol{\lambda}}^k)_q D_{v_{qi}} \right) \frac{\partial^2 \mathrm{f}_i^k}{\partial y_l \partial y_p} \end{bmatrix}. \tag{11}$$

The analysis is then simply as follows:

1) In principle, if we start with a well-defined initial estimate $\mathbf{f}_{\boldsymbol{\lambda}}^0$, the algorithm described by equations (2)-(4) is designed so that $\mathbf{M}^k$ and $\mathbf{f}_{\boldsymbol{\lambda}}^{k+1}$ are well-defined for all $k \geq 1$. Moreover, $\mathbf{M}^k$ is a full-rank matrix and therefore has a stable inverse $(\mathbf{M}^k)^{-1}$.
2) It should be noted all the elements of $\mathbf{M}^k$ are differentiable because (4) is a true function of $\mathbf{y}$. Thus, all the derivatives involved in this analysis are in the true sense of differentiation and not in the weak sense of distributions.
3) Then, (8) and (9) ensure that $\frac{\partial M_{mn}^k}{\partial y_p}$ is well-defined $\forall\ m,n,p = 1,2,\ldots,N$, provided Λ^k is well-conditioned which is the case as long as $w_q^k < +\infty$, $\forall\ q = 1,2,\ldots,N$, and $k \geq 1$. This

can be ensured numerically. Therefore, for a fixed p, a well-defined $N \times 1$ vector $\frac{\partial \mathbf{f}_{\boldsymbol{\lambda}}^{k+1}}{\partial y_p}$ is obtained from (6) as

$$\frac{\partial \mathbf{f}_{\boldsymbol{\lambda}}^{k+1}}{\partial y_p} = (\mathbf{M}^k)^{-1}(\mathbf{e}_p - \mathbf{S}_p^k \mathbf{f}_{\boldsymbol{\lambda}}^{k+1}), \tag{12}$$

where $\mathbf{S}_p^k$ is a $N \times N$ matrix such that $(\mathbf{S}_p^k)_{mn} = \frac{\partial M_{mn}^k}{\partial y_p}$, $\mathbf{e}_p$ is a $N \times 1$ column vector whose elements are all zeros except the p^{th} one which is unity.

4) While (10) and (11) ensure that $\frac{\partial^2 M_{mn}^k}{\partial y_l \partial y_p}$ is well-defined $\forall\ m, n, l, p = 1, 2, \ldots, N$, equations (8) and (12) ensure that the second and third terms in the L.H.S of (7) are well-defined. Thus, we see that for a given (l, p) a well-defined $N \times 1$ vector $\frac{\partial \mathbf{f}_{\boldsymbol{\lambda}}{}^{k+1}}{\partial y_l \partial y_p}$ is obtained from (7) as

$$\frac{\partial \mathbf{f}_{\boldsymbol{\lambda}}{}^{k+1}}{\partial y_l \partial y_p} = -(\mathbf{M}^k)^{-1}\left(\mathbf{P}_{lp}^k \mathbf{f}_{\boldsymbol{\lambda}}^{k+1} + \mathbf{S}_p^k \frac{\partial \mathbf{f}_{\boldsymbol{\lambda}}^{k+1}}{\partial y_p} + \mathbf{S}_l^k \frac{\partial \mathbf{f}_{\boldsymbol{\lambda}}^{k+1}}{\partial y_p}\right), \tag{13}$$

where $\mathbf{P}_{lp}^k$ is a $N \times N$ matrix such that $(\mathbf{P}_{lp}^k)_{mn} = \frac{\partial^2 M_{mn}^k}{\partial y_l \partial y_p}$. ■

II. Monte-Carlo Divergence Estimation under a Weaker Hypothesis

Here, we restate the second part of Theorem 2 which deals with the Monte-Carlo divergence estimation under the weaker hypothesis of tempered distributions and then give a formal proof of this result.

Theorem 2: Let $\mathbf{b}'$ be a zero-mean unit variance i.i.d random vector. Assume that $\exists\ n_0 > 1$ and $C_0 > 0$ such that

$$\|\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})\| \leq C_0(1 + \|\mathbf{y}\|^{n_0}), \tag{14}$$

that is, $\mathbf{f}_{\boldsymbol{\lambda}}$ is tempered. Then

$$\mathrm{div}_{\mathbf{y}}\{\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})\} = \lim_{\varepsilon \to 0} E_{\mathbf{b}'}\left\{\mathbf{b}'^{\mathrm{T}}\left(\frac{\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y} + \varepsilon \mathbf{b}') - \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})}{\varepsilon}\right)\right\} \tag{15}$$

in the weak-sense of tempered distributions.

Proof: Let $\psi \in \mathcal{S}$ be a rapidly decaying (test) function that is infinitely differentiable. We have to show that

$$\langle \mathrm{div}\{\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})\}, \psi(\mathbf{y})\rangle = \lim_{\varepsilon \to 0}\left\langle E_{\mathbf{b}'}\left\{\mathbf{b}'^{\mathrm{T}}\left(\frac{\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y} + \varepsilon \mathbf{b}') - \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})}{\varepsilon}\right)\right\}, \psi(\mathbf{y})\right\rangle. \tag{16}$$

We note that the L.H.S of (16) can be expressed as (from theory of distributions)

$$\langle \mathrm{div}\{\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})\}, \psi(\mathbf{y})\rangle = -\langle \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y}), \nabla \psi(\mathbf{y})\rangle. \tag{17}$$

The R.H.S of (16) involves the double integration

$$I_1(\varepsilon) = \int_{\mathbf{y}} d\mathbf{y}\, \psi(\mathbf{y}) \int_{\mathbf{b}'} \mathbf{b}'^{\mathrm{T}}\left(\frac{\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y} + \varepsilon \mathbf{b}') - \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})}{\varepsilon}\right) q(\mathbf{b}') d\mathbf{b}', \tag{18}$$

where $q(\mathbf{b}')$ represents the probability density function of $\mathbf{b}'$. The order of the integration can be swapped as soon as (Fubini's Theorem)

$$I_2(\varepsilon) = \int_{\mathbf{y}} \int_{\mathbf{b}'} |\psi(\mathbf{y})| \left|\mathbf{b}'^{\mathrm{T}}\left(\frac{\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y} + \varepsilon \mathbf{b}') - \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})}{\varepsilon}\right)\right| q(\mathbf{b}')\, d\mathbf{b}' d\mathbf{y} < +\infty. \tag{19}$$

Using Triangle inequality, we bound $I_2(\varepsilon)$ as

$$I_2(\varepsilon) \leq \frac{J(\varepsilon) + J(0)}{\varepsilon}, \tag{20}$$

where

$$J(\varepsilon) = \int_{\mathbf{y}} \int_{\mathbf{b}'} |\psi(\mathbf{y})| \left| \mathbf{b}'^{\mathrm{T}} \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y} + \varepsilon \mathbf{b}') \right| q(\mathbf{b}') \, d\mathbf{b}' d\mathbf{y}. \tag{21}$$

Using Cauchy-Schwarz inequality and the fact that $\mathbf{f}_{\boldsymbol{\lambda}}$ is tempered (cf. (14)), we get

$$J(\varepsilon) \leq \int_{\mathbf{y}} \int_{\mathbf{b}'} |\psi(\mathbf{y})| \, \|\mathbf{b}'\| \, C_0 (1 + \|\mathbf{y} + \varepsilon \mathbf{b}'\|^{n_0}) \, q(\mathbf{b}') \, d\mathbf{b}' d\mathbf{y}. \tag{22}$$

Using the convexity property of the function $(\cdot)^{n_0}$ for $n_0 > 1$, we get

$$\begin{aligned} J(\varepsilon) &\leq \int_{\mathbf{y}} \int_{\mathbf{b}'} |\psi(\mathbf{y})| \, \|\mathbf{b}'\| \, C_0 (1 + 2^{n_0 - 1} \|\mathbf{y}\|^{n_0} + 2^{n_0 - 1} \|\varepsilon \mathbf{b}'\|^{n_0}) \, q(\mathbf{b}') \, d\mathbf{b}' d\mathbf{y} \\ &= C_0 \left(\int_{\mathbf{y}} |\psi(\mathbf{y})| \, d\mathbf{y} \right) \left(\int_{\mathbf{b}'} \|\mathbf{b}'\| \, q(\mathbf{b}') \, d\mathbf{b}' \right) \\ &\quad + C_0 2^{n_0 - 1} \left(\int_{\mathbf{y}} |\psi(\mathbf{y})| \, \|\mathbf{y}\|^{n_0} \, d\mathbf{y} \right) \left(\int_{\mathbf{b}'} \|\mathbf{b}'\| \, q(\mathbf{b}') \, d\mathbf{b}' \right) \\ &\quad + C_0 2^{n_0 - 1} \varepsilon^{n_0} \left(\int_{\mathbf{y}} |\psi(\mathbf{y})| \, d\mathbf{y} \right) \left(\int_{\mathbf{b}'} \|\mathbf{b}'\|^{n_0 + 1} \, q(\mathbf{b}') \, d\mathbf{b}' \right) \\ &< +\infty, \end{aligned} \tag{23}$$

under the hypothesis that $E_{\mathbf{b}'}\{\|\mathbf{b}'\|^{n_0}\} < +\infty, \ \forall \ n_0 \geq 1$. The ones involving ψ are also finite because ψ is a rapidly decaying function with finite support. Thus, $J(\varepsilon) < \infty, \ \forall \ \varepsilon \geq 0$. Hence, we interchange the integrals (with appropriate change of variables) to get

$$\begin{aligned} I_1(\varepsilon) &= \int_{\mathbf{b}'} q(\mathbf{b}') d\mathbf{b}' \int_{\mathbf{y}} \psi(\mathbf{y}) \mathbf{b}'^{\mathrm{T}} \left(\frac{\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y} + \varepsilon \mathbf{b}') - \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})}{\varepsilon} \right) d\mathbf{y} \\ &= \int_{\mathbf{b}'} q(\mathbf{b}') d\mathbf{b}' \int_{\mathbf{y}} \mathbf{b}'^{\mathrm{T}} \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y}) \left(\frac{\psi(\mathbf{y} - \varepsilon \mathbf{b}') - \psi(\mathbf{y})}{\varepsilon} \right) d\mathbf{y}. \end{aligned} \tag{24}$$

Since ψ is infinitely differentiable, we apply Taylor's theorem [2] to $\psi(\mathbf{y} - \varepsilon \mathbf{b}')$ and obtain

$$\frac{\psi(\mathbf{y} - \varepsilon \mathbf{b}') - \psi(\mathbf{y})}{\varepsilon} = - \int_0^1 \mathbf{b}'^{\mathrm{T}} \, \nabla \psi(\mathbf{y} - t \, \varepsilon \mathbf{b}') \, dt. \tag{25}$$

Therefore,

$$I_1(\varepsilon) = - \int_{\mathbf{b}'} q(\mathbf{b}') d\mathbf{b}' \int_{\mathbf{y}} \mathbf{b}'^{\mathrm{T}} \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y}) \int_0^1 \mathbf{b}'^{\mathrm{T}} \, \nabla \psi(\mathbf{y} - t \, \varepsilon \mathbf{b}') \, dt \, d\mathbf{y}. \tag{26}$$

We want to let ε tend to 0 in the above expression. This is accomplished by the application of Lebesgue's dominated convergence theorem. But firstly, we must bound the integrand

$$z(\mathbf{y}, \mathbf{b}', t, \varepsilon) = -q(\mathbf{b}') \ \mathbf{b}'^{\mathrm{T}} \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y}) \ \mathbf{b}'^{\mathrm{T}} \nabla \psi(\mathbf{y} - t \, \varepsilon \mathbf{b}') \ \beta^0(t), \tag{27}$$

by an integrable function $g(\mathbf{y}, \mathbf{b}', t)$, where

$$\beta^0(t) = \begin{cases} 1, & \text{if } t \in (0, 1) \\ 0, & \text{otherwise} \end{cases}. \tag{28}$$

To do that, we start with $|z(\mathbf{y}, \mathbf{b}', t, \varepsilon)|$ and apply Cauchy-Schwarz inequality to obtain

$$0 \leq |z(\mathbf{y}, \mathbf{b}', t, \varepsilon)| \ \leq \ \underbrace{q(\mathbf{b}') \, \|\mathbf{b}'\|^2 \, \|\nabla \psi(\mathbf{y} - t \, \varepsilon \mathbf{b}')\| \, \beta^0(t)}_{\stackrel{\text{def}}{=} g_0(\mathbf{y}, \mathbf{b}', t, \varepsilon)} \, \|\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})\|. \tag{29}$$

Now, by using convexity of $(\cdot)^{n_1}$ for $n_1 \geq 1$, we have

$$1 + \|\mathbf{y}\|^{n_1} = 1 + \|\mathbf{y} - \varepsilon t \mathbf{b}' + \varepsilon t \mathbf{b}'\|^{n_1} \leq 1 + 2^{n_1 - 1} \|\mathbf{y} - \varepsilon t \mathbf{b}'\|^{n_1} + 2^{n_1 - 1} \|\varepsilon t \mathbf{b}'\|^{n_1}. \tag{30}$$

Then, for $\varepsilon \leq 1$,

$$\begin{aligned}(1+\|\mathbf{y}\|^{n_1})\, g_0(\mathbf{y},\mathbf{b}',t,\varepsilon) &\leq (1+2^{n_1-1}\|\mathbf{y}-\varepsilon t\mathbf{b}'\|^{n_1})\|\nabla\psi(\mathbf{y}-t\,\varepsilon\mathbf{b}')\|\, q(\mathbf{b}')\,\|\mathbf{b}'\|^2\beta^0(t) \\ &\quad +2^{n_1-1}\|\nabla\psi(\mathbf{y}-t\,\varepsilon\mathbf{b}')\|\, q(\mathbf{b}')\,\|\mathbf{b}'\|^{n_1+2}\, t^{n_1}\beta^0(t) \\ &\leq \underbrace{C_{\psi,1}\, q(\mathbf{b}')\,\|\mathbf{b}'\|^2\beta^0(t)+C_{\psi,2}\, q(\mathbf{b}')\,\|\mathbf{b}'\|^{n_1+2}\, t^{n_1}\beta^0(t)}_{\stackrel{\text{def}}{=} g_1(\mathbf{b}',t)},\end{aligned}$$

where $C_{\psi,1}=\sup_{\mathbf{y}}\{(1+2^{n_1-1}\|\mathbf{y}\|^{n_1})\,\|\nabla\psi(\mathbf{y})\|\}$ and $C_{\psi,2}=2^{n_1-1}\sup_{\mathbf{y}}\{\|\nabla\psi(\mathbf{y})\|\}$.

Since $E_{\mathbf{b}'}\{\|\mathbf{b}'\|^{n_1+2}\}<+\infty$, it is clear that

$$\int_{\mathbf{b}'}\int_t g_1(\mathbf{b}',t)\, d\mathbf{b}'\, dt<+\infty. \tag{31}$$

Therefore, choosing $n_1>n_0+N$, where N is the dimension of $\mathbf{y}$, and $\forall\ \varepsilon\leq 1$ we see that

$$\begin{aligned}|z(\mathbf{y},\mathbf{b}',t,\varepsilon)|\leq g_0(\mathbf{y},\mathbf{b}',t,\varepsilon)\,\|\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})\| &\leq g_1(\mathbf{b}',t)\frac{\|\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})\|}{1+\|\mathbf{y}\|^{n_1}} \\ &\leq \underbrace{C_0\, g_1(\mathbf{b}',t)\frac{1+\|\mathbf{y}\|^{n_0}}{1+\|\mathbf{y}\|^{n_1}}}_{\stackrel{\text{def}}{=} g(\mathbf{y},\mathbf{b}',t)}.\end{aligned} \tag{32}$$

Then, we notice that

$$\begin{aligned}\int_{\mathbf{y}}\frac{1+\|\mathbf{y}\|^{n_0}}{1+\|\mathbf{y}\|^{n_1}}d\mathbf{y} &= \sum_{\mathbf{k}\in\mathbb{Z}^N}\int_{[0,\ 1)^N}\frac{1+\|\mathbf{y}+\mathbf{k}\|^{n_0}}{1+\|\mathbf{y}+\mathbf{k}\|^{n_1}}d\mathbf{y}=\int_{[0,\ 1)^N}\left(\sum_{\mathbf{k}\in\mathbb{Z}^N}\frac{1+\|\mathbf{y}+\mathbf{k}\|^{n_0}}{1+\|\mathbf{y}+\mathbf{k}\|^{n_1}}\right)d\mathbf{y} \quad \text{(Fubini)} \\ &\leq \sum_{\mathbf{k}\in\mathbb{Z}^N}\frac{1+\|\mathbf{1}+\mathbf{k}\|^{n_0}}{1+\|\mathbf{k}\|^{n_1}} \quad (\mathbf{1}\text{ is a column vector of 1s}) \\ &\leq \sum_{\mathbf{k}\in\mathbb{Z}^N}\frac{1+2^{n_0-1}N^{\frac{n_0}{2}}+2^{n_0-1}\|\mathbf{k}\|^{n_0}}{1+\|\mathbf{k}\|^{n_1}} \quad \text{(using convexity of } (\cdot)^{n_0}) \\ &\leq (1+2^{n_0-1}N^{\frac{n_0}{2}})\left(1+\underbrace{\sum_{\mathbf{k}\in\mathbb{Z}^N\setminus\{\mathbf{0}\}}\frac{1}{\|\mathbf{k}\|^{n_1}}}_{<+\infty}\right)+2^{n_0-1}\underbrace{\sum_{\mathbf{k}\in\mathbb{Z}^N\setminus\{\mathbf{0}\}}\frac{1}{\|\mathbf{k}\|^{n_1-n_0}}}_{<+\infty} \\ &< +\infty,\end{aligned} \tag{33}$$

whenever $n_1>n_0+N$.

Because of (31) and (33), we find

$$\int_{\mathbf{y}}\int_{\mathbf{b}'}\int_t g(\mathbf{y},\mathbf{b}',t)\, d\mathbf{y}\, d\mathbf{b}'\, dt=\left(\int_{\mathbf{b}'}\int_t g(\mathbf{y},\mathbf{b}',t)\, d\mathbf{b}'\, dt\right)\left(\int_{\mathbf{y}}\frac{1+\|\mathbf{y}\|^{n_0}}{1+\|\mathbf{y}\|^{n_1}}d\mathbf{y}\right)<+\infty. \tag{34}$$

Therefore, z qualifies for both Fubini's and Lebesgue's Dominant Convergence Theorems (cf. (32)

and (34)). Hence, applying the limit with appropriate change of integrals, we get the desired result:

$$\begin{aligned}
\lim_{\varepsilon\to 0} I_1(\varepsilon) &= -\lim_{\varepsilon\to 0}\int_{\mathbf{b}'} q(\mathbf{b}')d\mathbf{b}' \int_{\mathbf{y}} \mathbf{b}'^{\mathrm{T}}\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y}) \int_0^1 \mathbf{b}'^{\mathrm{T}}\,\nabla\psi(\mathbf{y}-t\,\varepsilon\mathbf{b}')\,dt\,d\mathbf{y} \\
&= -\int_{\mathbf{b}'} q(\mathbf{b}')d\mathbf{b}' \int_{\mathbf{y}} \mathbf{b}'^{\mathrm{T}}\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y}) \int_0^1 \lim_{\varepsilon\to 0}\mathbf{b}'^{\mathrm{T}}\,\nabla\psi(\mathbf{y}-t\,\varepsilon\mathbf{b}')\,dt\,d\mathbf{y} \\
&= -\int_{\mathbf{y}} \mathbf{f}_{\boldsymbol{\lambda}}^{\mathrm{T}}(\mathbf{y}) \underbrace{\left(\int_{\mathbf{b}'} q(\mathbf{b}')\mathbf{b}'\mathbf{b}'^{\mathrm{T}} d\mathbf{b}'\right)}_{=I} \nabla\psi(\mathbf{y})\,d\mathbf{y} \\
&= -\langle \mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y}), \nabla\psi(\mathbf{y})\rangle = \langle \mathrm{div}\{\mathbf{f}_{\boldsymbol{\lambda}}(\mathbf{y})\}, \psi(\mathbf{y})\rangle \quad \text{(from (17))}\ . \quad \blacksquare
\end{aligned}$$

References

[1] M. A. T. Figueiredo, J. B. Dias, J. P. Oliveira, and R. D. Nowak, "On total variation denoising: A new Majorization-Minimization algorithm and an experimental comparison with wavalet denoising," *Proceedings of IEEE International Conference on Image Processing (ICIP 2006), Atlanta, GA, USA*, pp. 2633–2636, October 2006.

[2] J. T. Day, "On the convergence of Taylor series for functions of n variables", *Mathematics Magazine*, vol. 40, No. 5, pp. 258–260, 1967.