A. MNIST-like Experiments

The plots presented in this appendix support the conclusions made in the main text and provide an overview of the experiments conducted on MNIST-like datasets.

Each figure shows the compression-accuracy trade-off of a particular method and input features for *SimpleConvModel* and *TwoLayerDenseModel* models for all four of the studied datasets (described in the main text): EMNIST-Letters on the *top-left*, KMNIST – *top-right*, Fashion MNIST – *bottom-left*, and MNIST on the *bottom-right*. Figures 7, 8, 9, and 10 present $\mathbb{R}$ and $\mathbb{C}$ models with *the same intermediate feature sizes*.

We compare $\mathbb{R}$ networks against $\frac{1}{2}\mathbb{C}$ with half the number of parameters for raw input features on figures 13, and 14, and $2\mathbb{R}$ with double the number of parameters against $\mathbb{C}$ for Fourier input features on figures 11 and 12.

B. Complex-valued Local Reparameterization

In this section we show (11).

By $e_i \in \mathbb{R} \hookrightarrow \mathbb{C}$ we denote the i-th unit vector of dimensionality *conforming* to the matrix-vector expression it is used in, $[M]$ denotes *row-major* flattening of a matrix M into a vector, i.e. in lexicographic order of its indices. Furthermore $\operatorname{diag}(\cdot)$ embeds vectors into matrices with zeros everywhere except the diagonal, and $\otimes$ is the Kronecker product, for which we note the following identities $[PQR] = (P \otimes R^\top)[Q]$, $(P \otimes Q)^\top = (P^\top \otimes Q^\top)$, and $(P \otimes R)(C \otimes S) = PQ \otimes RS$ (Petersen and Pedersen, 2012).

If we assume a factorized $\mathbb{C}$-Gaussian approximation (10) for $W \in \mathbb{C}^{n\times m}$, then $[W]$ is $\mathbb{C}$-Gaussian vector with

$$[W] \sim \mathcal{CN}_{nm}\bigl([\mu], \operatorname{diag}[\Sigma], \operatorname{diag}[C]\bigr)\,, \tag{17}$$

where with $C_{ij} = \Sigma_{ij}\xi_{ij}$, $\Sigma_{ij} \geq 0$, and $|C_{ij}|^2 \leq \Sigma_{ij}$. Then for any $x \in \mathbb{C}^m$ and $b \in \mathbb{C}^n$ we have $y = Wx + b = (I_n \otimes x^\top)[W] + b$, whence the covariance and relation matrices of y are

$$\begin{aligned}
\bigl(I_n \otimes x^\top\bigr)\operatorname{diag}[\Sigma]\overline{\bigl(I_n \otimes x^\top\bigr)}^\top &= \sum_{ij}\bigl(I_n \otimes x^\top\bigr)\Bigl((e_i \otimes e_j)\Sigma_{ij}(e_i \otimes e_j)^\top\Bigr)\overline{\bigl(I_n \otimes x^\top\bigr)}^\top \\
&= \sum_{ij}\bigl(e_i \otimes x^\top e_j\bigr)\Sigma_{ij}\overline{\bigl(e_i \otimes x^\top e_j\bigr)}^\top \\
&= \sum_{i=1}^{n}\bigl(e_i e_i^\top\bigr)\Biggl\{\sum_{j=1}^{m}\Sigma_{ij}|x_j|^2\Biggr\}\,,
\end{aligned} \tag{18}$$

$$\begin{aligned}
\bigl(I_n \otimes x^\top\bigr)\operatorname{diag}[C]\bigl(I_n \otimes x^\top\bigr)^\top &= \sum_{ij}\bigl(I_n \otimes x^\top\bigr)\Bigl((e_i \otimes e_j)C_{ij}(e_i \otimes e_j)^\top\Bigr)\bigl(I_n \otimes x^\top\bigr)^\top \\
&= \sum_{i=1}^{n}\bigl(e_i e_i^\top\bigr)\Biggl\{\sum_{j=1}^{m}C_{ij}x_j^2\Biggr\}\,.
\end{aligned} \tag{19}$$

Since (18) and (19) are diagonal, the vector y has independent univariate $\mathbb{C}$-Gaussian components, whence (11) follows.

C. Backpropagation through $\mathbb{C}$-networks

Wirtinger ($\mathbb{CR}$) calculus relies on the natural identification of $\mathbb{C}$ with $\mathbb{R}^2$, and regards $f\colon \mathbb{C} \to \mathbb{C}$ as an algebraically equivalent function $F\colon \mathbb{R}^2 \to \mathbb{C}$ defined $f(z) = f(u + \jmath v) = F(u, v)$. It enables general treatment of functions of vector $\mathbb{C}$-argument that possess partial derivatives with respect to real and imaginary parts, yet are not required to satisfy Cauchy-Riemann conditions. In $\mathbb{CR}$ calculus the complex argument z and its conjugate $\overline{z}$ act as independent variables and $f(z)$ is treated as $f(z, \overline{z})$ by way of geometric transformations $z = u + \jmath v$ and $\overline{z} = u - \jmath v$.

Wirtinger partial derivative operators are formally defined as $\frac{\partial}{\partial z} = \frac{1}{2}\bigl(\frac{\partial}{\partial u} - \jmath\frac{\partial}{\partial v}\bigr)$ and $\frac{\partial}{\partial \overline{z}} = \frac{1}{2}\bigl(\frac{\partial}{\partial u} + \jmath\frac{\partial}{\partial v}\bigr)$ and differentials are $dz = du + \jmath dv$ and $d\overline{z} = du - \jmath dv$. In this paradigm The usual rules of calculus, like chain and product rules, follow directly from the definition of the operators, e.g.

$$\frac{\partial (f \circ g)}{\partial z} = \frac{\partial f(g(z))}{\partial g} \frac{\partial g(z)}{\partial z} + \frac{\partial f(g(z))}{\partial \overline{g}} \frac{\partial \overline{g(z)}}{\partial z} \,.$$

The total differential of f at $z = u + \jmath v \in \mathbb{C}$ is

$$\begin{aligned} df(z) &= \frac{\partial f}{\partial z} dz + \frac{\partial f}{\partial \overline{z}} d\overline{z} \\ &= \frac{1}{2}\left(\frac{\partial F}{\partial u} du - \jmath \cancel{\frac{\partial F}{\partial v}} du + \jmath \cancel{\frac{\partial F}{\partial u}} dv + \frac{\partial F}{\partial v} dv \right) + \frac{1}{2}\left(\frac{\partial F}{\partial u} du + \jmath \cancel{\frac{\partial F}{\partial v}} du - \jmath \cancel{\frac{\partial F}{\partial u}} dv + \frac{\partial F}{\partial v} dv \right) \\ &= dF(u, v) \,, \end{aligned}$$

At the same time the Cauchy-Riemann conditions $-\jmath \frac{\partial F}{\partial v} = \frac{\partial F}{\partial u}$ can be expressed as $\frac{\partial f}{\partial \overline{z}} = 0$. Thus $\mathbb{CR}$ calculus subsumes the usual $\mathbb{C}$-calculus of holomorphic functions, since $f(z) = f(z, \overline{z})$ is constant with respect to $\overline{z}$ in the latter.

In optimization-related tasks the objective is $f\colon \mathbb{C} \to \mathbb{R}$, meaning that if it were to satisfy the Cauchy-Riemann conditions, then it necessarily should have been constant. Nevertheless, the expression of the $\mathbb{CR}$ gradient is compatible with what is expected, when f is treated like a $\mathbb{R}^2$ function. For such f we have $\overline{f} = f$, which implies $\frac{\partial f}{\partial \overline{z}} = \frac{\partial \overline{f}}{\partial \overline{z}} = \overline{\frac{\partial f}{\partial z}}$, whence

$$df = \tfrac{\partial f}{\partial z} dz + \tfrac{\partial f}{\partial \overline{z}} d\overline{z} = \tfrac{\partial f}{\partial z} dz + \overline{\tfrac{\partial f}{\partial z} dz} = 2\Re\big(\tfrac{\partial f}{\partial z} dz\big) \,.$$

Therefore the gradient of f at z is given by $\nabla_{\overline{z}} f(z) = \overline{\frac{\partial f}{\partial z}} = \frac{\partial F}{\partial u} + \jmath \frac{\partial F}{\partial v}$. The identification $\mathbb{C} \simeq \mathbb{R}^2$, backed by Wirtinger calculus, and emulation of $\mathbb{C}$-arithmetic in computational graphs with $\mathbb{R}$-valued operations makes it possible to reuse $\mathbb{R}$ back-propagation and existing auto-differentiation frameworks.

D. Gradient of the KL-divergence in $\mathbb{R}$ case

In this appendix we study the approximation proposed by Molchanov et al. (2017) for the KL divergence term (4) for $\mathbb{R}$ Sparse Variational Dropout. Following the logic of Lapidoth and Moser (2003) we derive the expression for $\frac{d}{d \log \alpha} K(\alpha)$. Acknowledging that the same result was obtained by Hron et al. (2018, eq. (5)), we provide this appendix for the sake of completeness.

For $(z_i)_{i=1}^m \sim \mathcal{N}(0, 1)$ iid and $(\mu_i)_{i=1}^m \in \mathbb{R}$, the random variable $W = \sum_i (\mu_i + z_i)^2$ has non-central χ^2 distribution with shape m and non-centrality parameter $\lambda = \sum_i \mu_i^2$, i.e. $W \sim \chi^2_m(\lambda)$. Therefore, the divergence (4) has the form

$$K(\alpha) \propto \frac{1}{2} \mathbb{E}_{W \sim \chi^2_1\left(\frac{1}{\alpha}\right)} \log W \,. \tag{4'}$$

W can alternatively be represented as a Poisson mixture of ordinary χ^2 distributions: if $Z_{|J} \sim \chi^2_{m+2J}$ for $J \sim \mathcal{P}ois(\frac{\lambda}{2})$ then $W \sim Z$. Therefore, expanding the conditional expectation gives

$$\mathbb{E}_{W \sim \chi^2_m(\lambda)} \log W = \mathbb{E}\Big(\mathbb{E}\big(\log W \mid J\big)\Big) = \mathbb{E}_{J \sim \mathcal{P}ois(\frac{\lambda}{2})} \Big(\mathbb{E}_{W \sim \chi^2_{m+2J}} \log W \Big) \,. \tag{20}$$

Since χ^2_ν is Gamma distribution $\Gamma(\frac{\nu}{2}, \frac{1}{2})$, it can be shown that the logarithmic moment $\mathbb{E}_{W \sim \chi^2_\nu} \log W$ is $\psi\big(\frac{\nu}{2}\big) - \log \frac{1}{2}$, where ψ is the digamma function ($\psi(x) = \frac{d}{dx} \log \Gamma(x)$). By expanding expectation of a Poisson random variable we get $\mathbb{E}_{W \sim \chi^2_m(\lambda)} \log W = \log 2 + g_m\big(\frac{\lambda}{2}\big)$, where

$$g_m(x) = e^{-x} \sum_{j \geq 0} \frac{x^j}{j!} \psi\big(\tfrac{m+2J}{2}\big) \,. \tag{21}$$

Making use of the property $\psi(z+1) = \psi(z) + \frac{1}{z}$ of the digamma funciton for $z > 0$, we conclude that the power series in (21) converges for any $x \geq 0$. Therefore the derivative of (21) is given by

$$\frac{d}{dx} g_m(x) = -g_m(x) + e^{-x} \sum_{j \geq 0} \frac{x^j}{j!} \left(\psi\big(\tfrac{m+2j}{2}\big) + \frac{2}{m+2j} \right) \,. \tag{22}$$

By manipulating the partial sums within (22) we get

$$\frac{d}{dx} g_m(x) = e^{-x} \sum_{j \geq 0} \frac{x^j}{j!} \frac{1}{j + \frac{m}{2}} = e^{-x} x^{-\frac{m}{2}} \sum_{j \geq 0} \frac{1}{j!} \int_0^x t^{j + \frac{m}{2} - 1} dt \,. \tag{23}$$

Furthermore, the functions $t \mapsto \sum_{j=0}^{J} \frac{1}{j!} t^{j + \frac{m}{2} - 1}$ are non-decreasing on $(0, x)$ with growing J and converge to $t^{\frac{m}{2} - 1} e^t$, which implies by the Monotone Convergence Theorem that

$$\frac{d}{dx} g_m(x) = e^{-x} x^{-\frac{m}{2}} \int_0^x \sum_{j \geq 0} \frac{1}{j!} t^{j + \frac{m}{2} - 1} dt = e^{-x} x^{-\frac{m}{2}} \int_0^x t^{\left(\frac{m}{2} - 1\right)} e^t dt \,. \tag{24}$$

Substituting $u^2 = t$ on $[0, \infty]$ with $2udu = dt$ and letting $I_m \colon x \mapsto e^{-x^2} \int_0^x u^{m-1} e^{u^2} du$ yields

$$\frac{d(20)}{d\lambda} = \frac{1}{2} \frac{d}{dx} g_m(x) \Big|_{x = \frac{\lambda}{2}} = e^{-x} x^{-\frac{m}{2}} \int_0^{\sqrt{x}} u^{m-1} e^{u^2} du \Big|_{x = \frac{\lambda}{2}} = \left(\sqrt{\tfrac{2}{\lambda}} \right)^m I_m\left(\sqrt{\tfrac{\lambda}{2}} \right) . \tag{25}$$

Since α is non-negative, it is typically parameterized via its logarithm, whence the derivative of (4') with respect to $\log \alpha$ follows from (25) for $m = 1$ and $\lambda = \frac{1}{\alpha}$:

$$\frac{dK(\alpha)}{d \log \alpha} = -\frac{1}{\sqrt{2\alpha}} I_1\left(\tfrac{1}{\sqrt{2\alpha}}\right) . \tag{26}$$

We compute the Monte-Carlo estimate of (4) on a sample of 10^7 draws over an equally spaced grid of $\log \alpha$ in $[-12, +12]$ of size 4096. The approximation proposed by Molchanov et al. (2017) is given in (27), with coefficients $k_1 = 0.63576$, $k_2 = 1.8732$, and $k_3 = 1.48695$. The derivative of the approximation with respect to $\log \alpha$ follows (26) within 4% of relative tolerance, see fig. 15.

$$(4) \approx \frac{1}{2} \log\left(1 + e^{-\log \alpha}\right) + k_1 \sigma\left(-(k_2 + k_3 \log \alpha)\right) , \tag{27}$$

Similarly, the forward difference estimate of the derivative (26) very closely (up to sampling error). For sake of completeness, we compute a similar Monte-Carlo estimate for the KL divergence term in (13') for $\mathbb{C}$-valued Variational Dropout with $\beta = 2$, fit the best approximation (27), and compare it against the exact $\log \alpha$ derivative $\frac{d(13')}{d \log \alpha} = e^{-\frac{1}{\alpha}} - 1$.

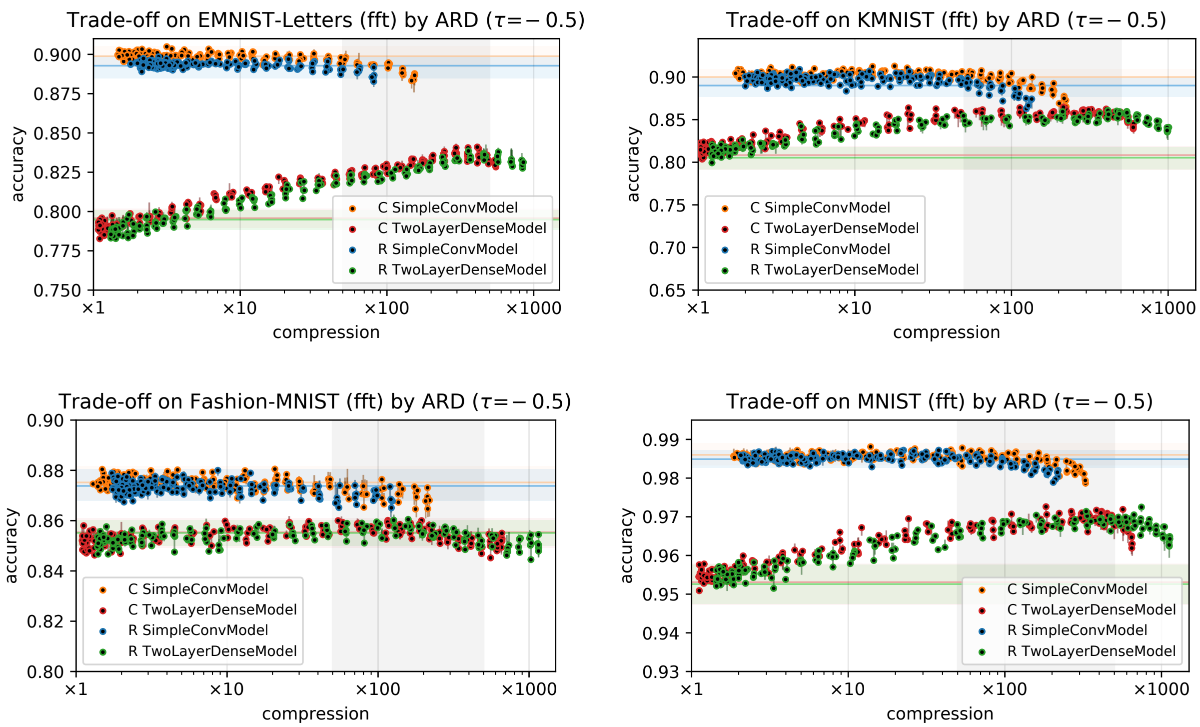

Figure 7. The trade-off of ARD method for ℝ and ℂ models using Fourier features.

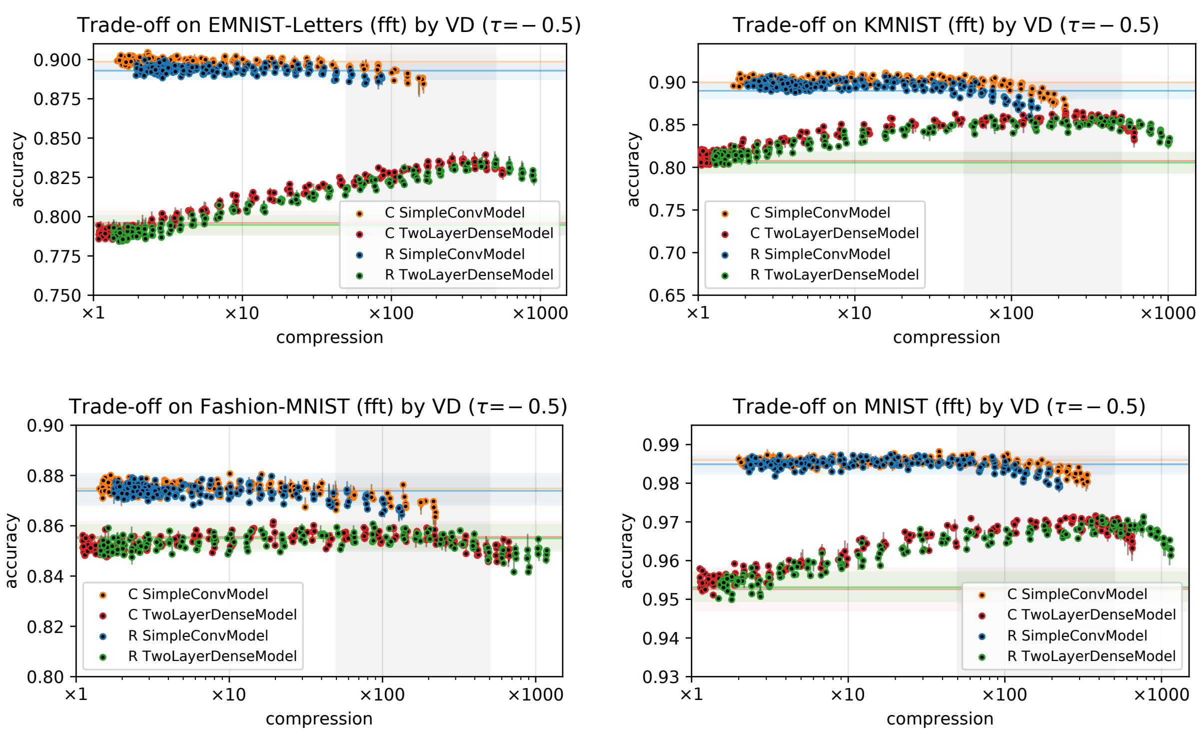

Figure 8. The trade-off of VD method for ℝ and ℂ models using Fourier features.

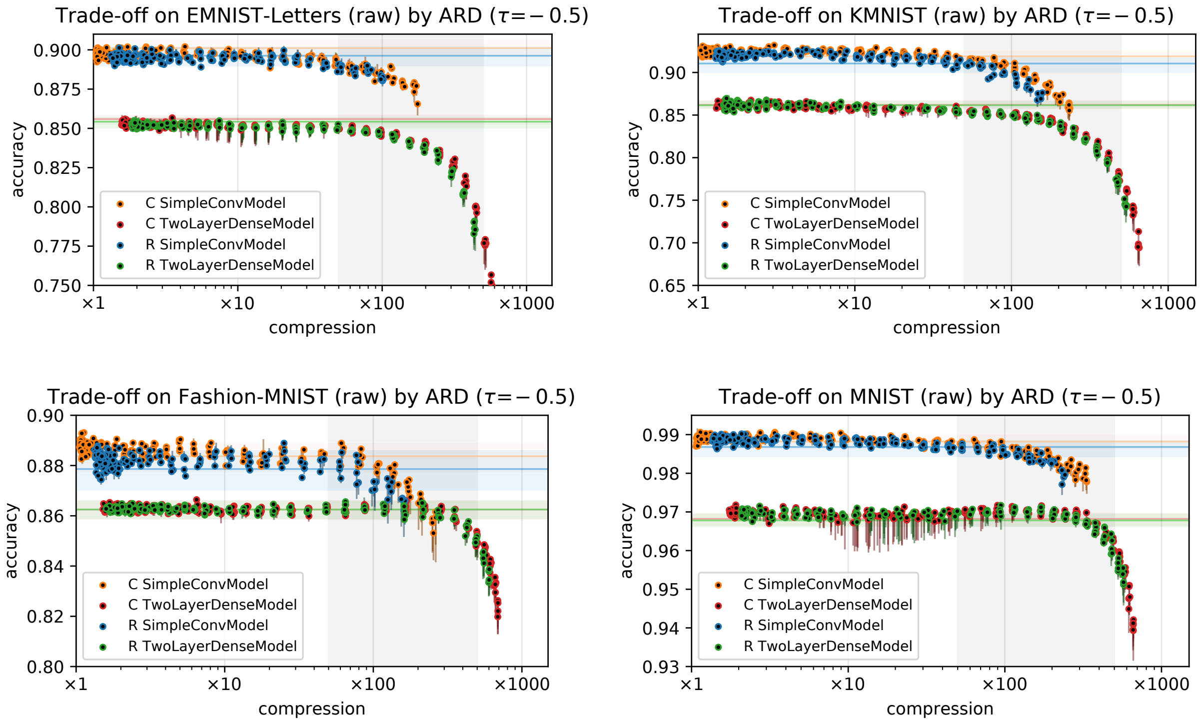

Figure 9. The trade-off of ARD method for ℝ and ℂ models using raw features.

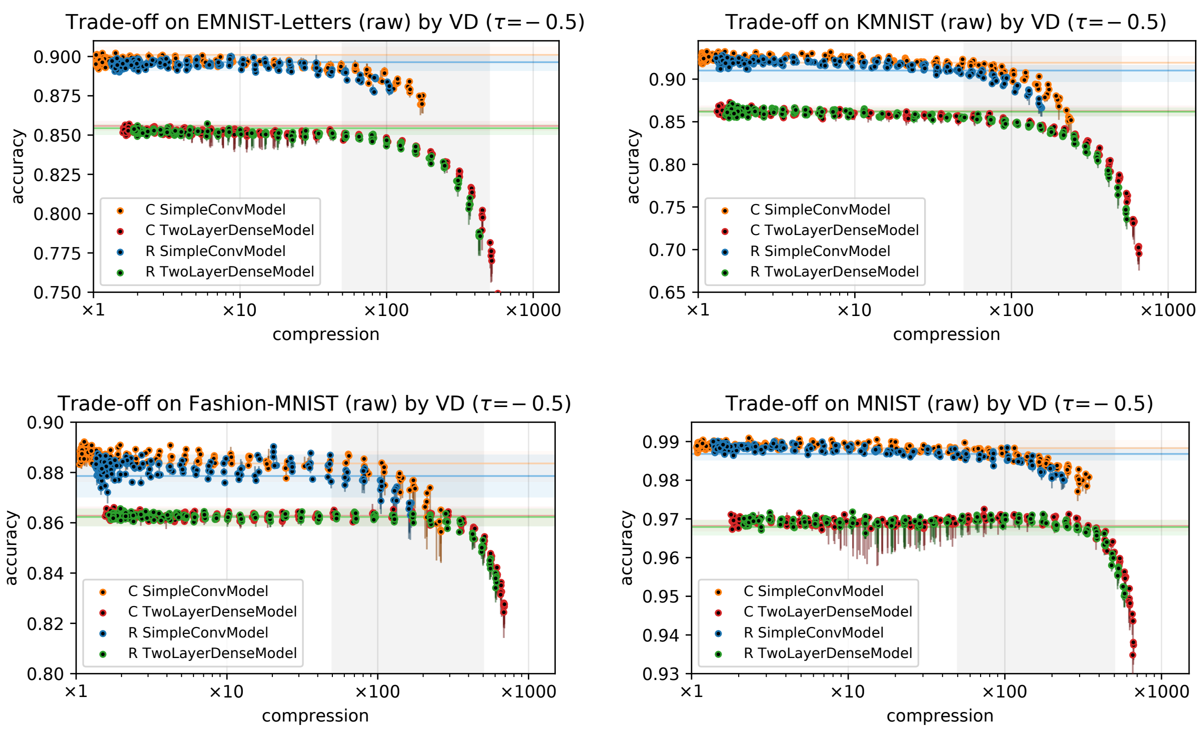

Figure 10. The trade-off of VD method for ℝ and ℂ models using raw features.

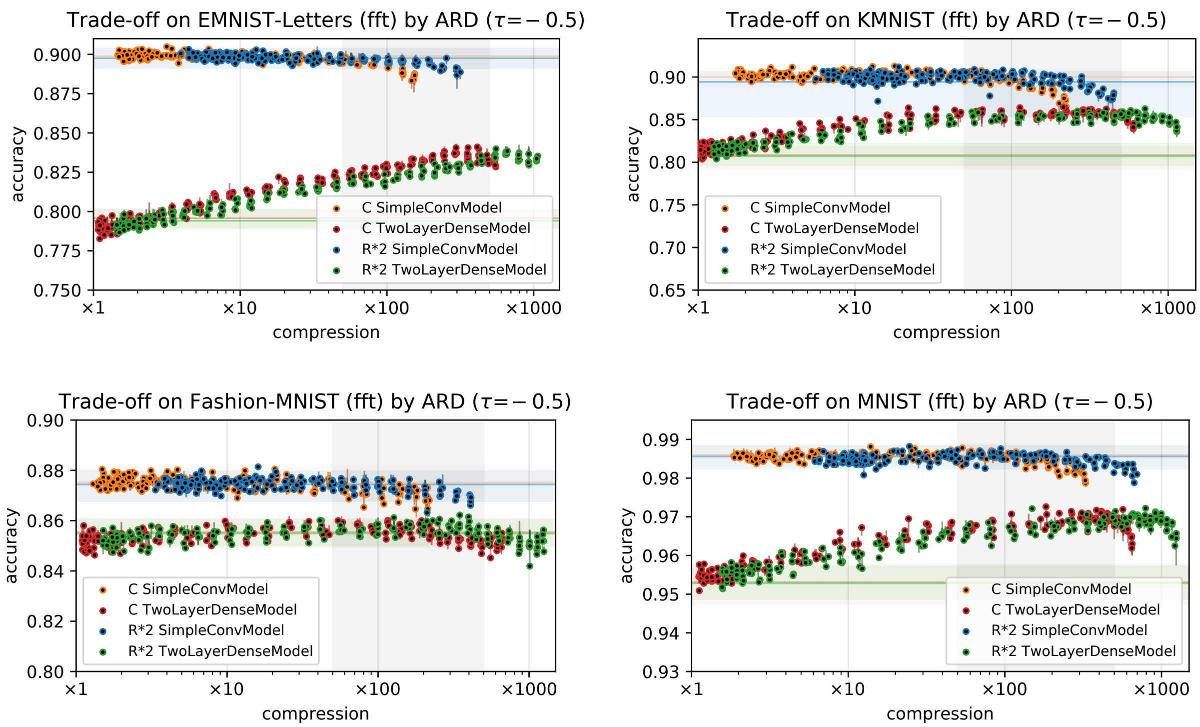

Figure 11. The trade-off of ARD method for $2\mathbb{R}$ and $\mathbb{C}$ models using Fourier features.

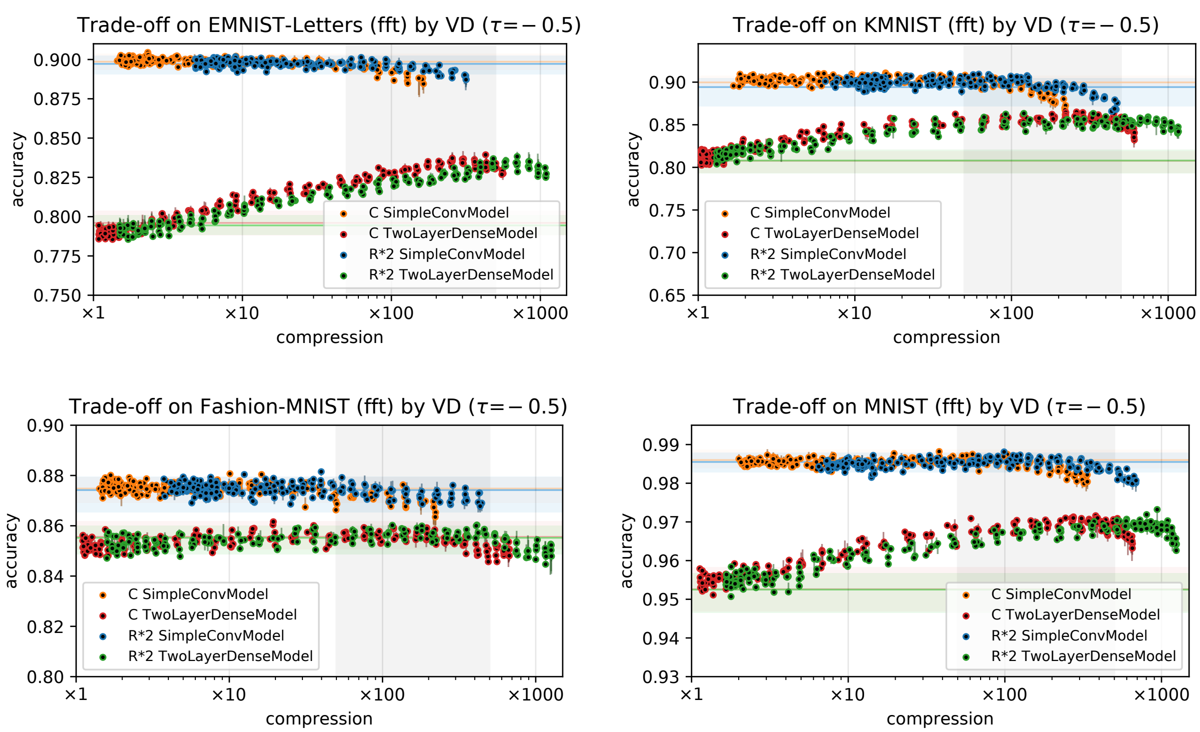

Figure 12. The trade-off of VD method for $2\mathbb{R}$ and $\mathbb{C}$ models using Fourier features.

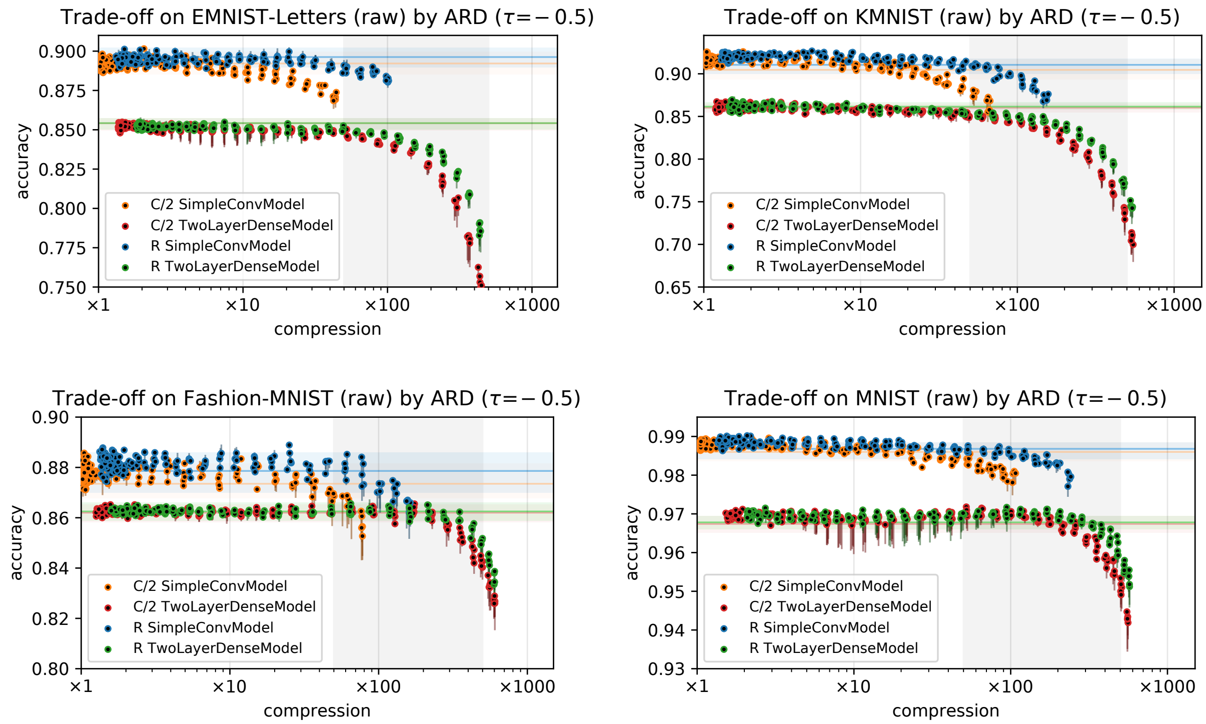

Figure 13. The trade-off of ARD method for $\mathbb{R}$ and $\frac{1}{2}\mathbb{C}$ models using raw features.

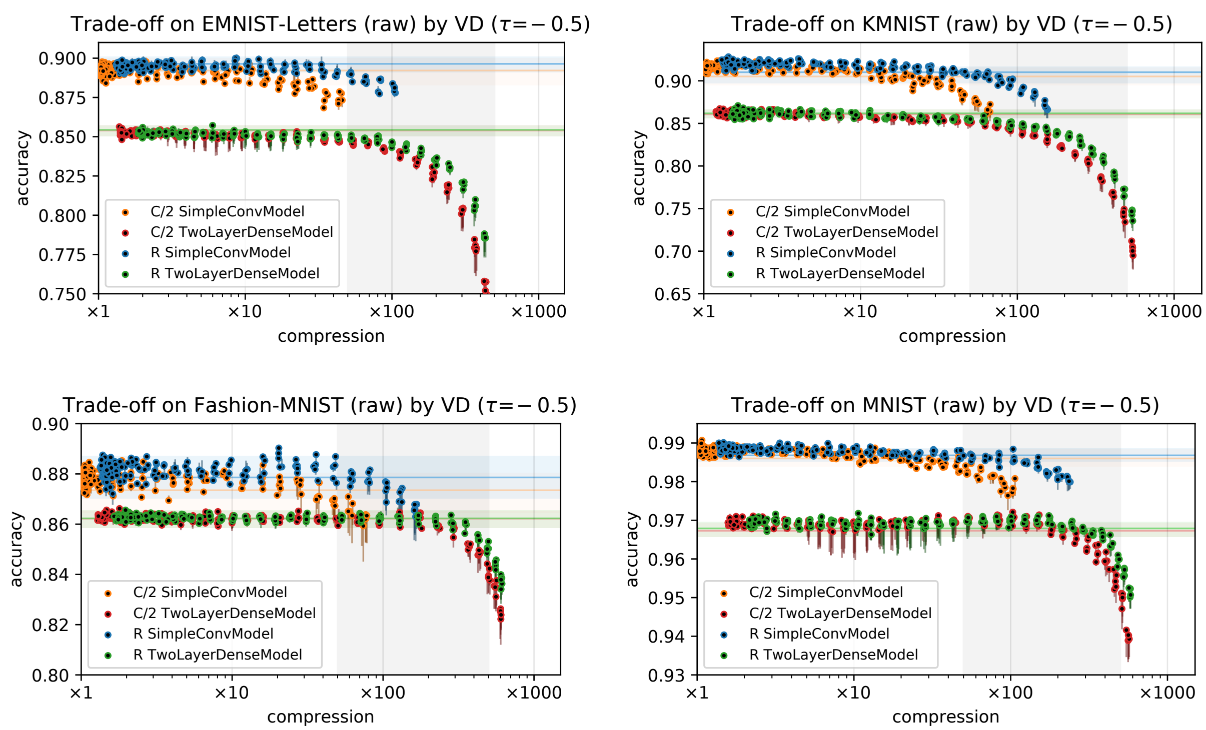

Figure 14. The trade-off of VD method for $\mathbb{R}$ and $\frac{1}{2}\mathbb{C}$ models using raw features.

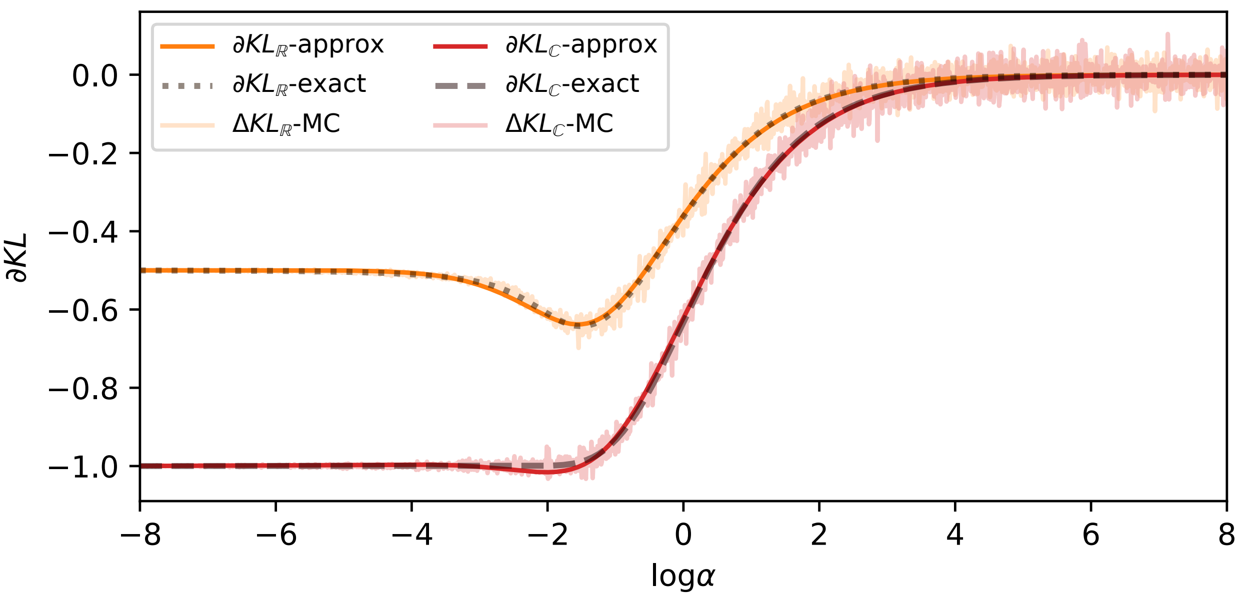

Figure 15. $\frac{dK(\;)}{d\log}$ of the approximation (27), MC estimate of (4), and the exact derivative using (26).