Backpropagation Beyond the Gradient

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Felix Julius Dangel, M. Sc.
aus Stuttgart

Tübingen
2022

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	To be announced
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Philipp Hennig
2. Berichterstatter:	Dr. Georg Martius

Disclaimer: this thesis uses Federico Marotta's `kaobook` template based on Ken Arroyo Ohori's doctoral thesis. I am grateful to Frank Schneider who added many stylistic features and would like to thank him, Katharina Ott, Frederik Kunstner, Julia Grosse, Jonathan Wenger, Matthias Werner, Marvin Pförtner, Runa Eschenhagen, Nathanel Bosch, and Andrés Fernández Rodríguez for their help to improve the text.

Acknowledgments

Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum. Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum. Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

Thank you!

Felix Dangel
Tübingen, August 31, 2022

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Zusammenfassung

Das hier ist der zweite Absatz. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext" oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muß keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum" dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Und nun folgt – ob man es glaubt oder nicht – der dritte Absatz. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext" oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muß keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum" dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Nach diesem vierten Absatz beginnen wir eine neue Zählung. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext" oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muß keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum" dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext" oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muß keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum" dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Das hier ist der zweite Absatz. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext" oder „Huardest gefburn"? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muß keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum" dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Table of Contents

Notation

The notation is influenced by Goodfellow et al. [2].

Tensors, Matrices, Vectors, Numbers

a	A scalar
$\boldsymbol{a}$	A column vector
$\boldsymbol{A}$	A matrix
A	A tensor
a_i or $[\boldsymbol{a}]_i$	The ith entry of the vector $\boldsymbol{a}$
$A_{i,j}$ or $[\boldsymbol{A}]_{i,j}$	The (i,j)th entry of the matrix $\boldsymbol{A}$ (row i, column j)
$[\boldsymbol{A}]_{i,:}$ (or $[\boldsymbol{A}]_{:,j}$)	The ith row (or jth column) of the matrix $\boldsymbol{A}$
$\mathsf{A}_{i,j,k}$ or $[\mathsf{A}]_{i,j,k}$	The (i,j,k)th entry of the tensor A
$\mathrm{vec}(\boldsymbol{A}), \mathrm{vec}(\mathsf{A})$	Matrix/tensor flattened into a vector; convention implies $\mathrm{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^\top \otimes \boldsymbol{A})\,\mathrm{vec}(\boldsymbol{B})$
$\mathrm{diag}(\boldsymbol{a})$	The square matrix with vector $\boldsymbol{a}$ on the diagonal and zeros elsewhere
$\mathrm{diag}(\boldsymbol{A})$	The vector containing the diagonal elements of the matrix $\boldsymbol{A}$
$\mathrm{diag}(\boldsymbol{A}_1, \dots, \boldsymbol{A}_L)$	A block-diagonal matrix with diagonal blocks given by square matrices $\boldsymbol{A}_1, \dots, \boldsymbol{A}_L$
$\mathrm{Tr}(\boldsymbol{A}), \det(\boldsymbol{A})$	Trace and determinant of a matrix $\boldsymbol{A}$
$\lVert\boldsymbol{a}\rVert_2$	L_2 norm of vector $\boldsymbol{a}$, *i.e.* $\lVert\boldsymbol{a}\rVert_2^2 = \boldsymbol{a}^\top\boldsymbol{a}$
$\mathrm{eig}(\boldsymbol{A}) := \{(\lambda_k, \boldsymbol{e}_k)\}_k$	Eigendecomposition of the matrix $\boldsymbol{A}$, eigenpairs $(\lambda_k, \boldsymbol{e}_k)$ satisfy $\boldsymbol{A}\boldsymbol{e}_k = \lambda_k \boldsymbol{e}_k$
$(\lambda_k(\boldsymbol{A}), \boldsymbol{e}_k(\boldsymbol{A}))$	kth eigenpair (eigenvalue, eigenvector) of matrix $\boldsymbol{A}$
$\boldsymbol{A} \otimes \boldsymbol{B}$	Kronecker product of two matrices, For two vectors $\boldsymbol{a}, \boldsymbol{b}$, one has $\boldsymbol{a} \otimes \boldsymbol{b}^\top = \boldsymbol{a}\boldsymbol{b}^\top$
$\mathsf{A} \odot \mathsf{B}, \boldsymbol{A} \odot \boldsymbol{B}, \boldsymbol{a} \odot \boldsymbol{b}$	Elementwise multiplication (Hadamard product) of two tensors, matrices, vectors
$\mathsf{A} \oslash \mathsf{B}, \boldsymbol{A} \oslash \boldsymbol{B}, \boldsymbol{a} \oslash \boldsymbol{b}$	Elementwise division (Hadamard division) of two tensors, matrices, vectors
$\mathsf{A}^{\odot 2}, \boldsymbol{A}^{\odot 2}, \boldsymbol{a}^{\odot 2}$	Elementwise square of a tensor, matrix, vector
$\mathsf{A}^{\odot 1/2}, \boldsymbol{A}^{\odot 1/2}, \boldsymbol{a}^{\odot 1/2}$	Elementwise square root of a tensor, matrix, vector

Empirical Risk Minimization

A datum is usually indicated by a subscript $_n$.

$(\boldsymbol{x}, \boldsymbol{y})$	Labeled datum with input features $\boldsymbol{x}$ and target $\boldsymbol{y}$
$(\boldsymbol{x}_n, \boldsymbol{y}_n)$	Datum n from a dataset
D	Total number of parameters in a model
$\boldsymbol{\theta} \in \Theta := \mathbb{R}^D$	Parameter vector of a model
$\boldsymbol{f} := f_{\boldsymbol{\theta}}(\boldsymbol{x})$	Prediction of a model $f_{\boldsymbol{\theta}}$ for input features $\boldsymbol{x}$
ℓ or $\ell(\boldsymbol{f}, \boldsymbol{y})$	Loss function to compare prediction and target; convex in $\boldsymbol{f}$
$\mathbb{D} := \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^{\lvert\mathbb{D}\rvert}$	A dataset containing instances of labeled data $(\boldsymbol{x}_n, \boldsymbol{y}_n)$ indexed by n
$\mathbb{B}$	A mini-batch $\mathbb{B} \subseteq \mathbb{D}$
N	Number of data in a mini-batch or a dataset, depending on the context
$\boldsymbol{f}_n := f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)$	Model prediction for datum n
ℓ_n or $\ell(\boldsymbol{f}_n, \boldsymbol{y}_n)$	Loss of datum n
$p_{\mathbb{D}}(\boldsymbol{x}, \boldsymbol{y})$	Empirical distribution of a dataset $\mathbb{D}$
$\mathcal{L}_{\mathbb{D}}(\boldsymbol{\theta})$	Empirical risk implied by the empirical distribution of a dataset $\mathbb{D}$
$\mathcal{L}_{\mathbb{D}_{\text{train}}}(\boldsymbol{\theta}), \mathcal{L}_{\mathbb{B}}(\boldsymbol{\theta})$, *etc.*	Training loss, mini-batch loss, *etc.*

Neural Networks

The layer number is indicated by parenthesized superscripts $^{(l)}$.

L	Total number of layers
$d^{(l)}$	Number of parameters in layer l; total number of parameters is $D = \sum_{l=1}^{L} d^{(l)}$
$\boldsymbol{\theta}^{(l)} \in \mathbb{R}^{d^{(l)}}$	Parameter vector of layer l, potentially empty for parameter-free layers like activations
$h^{(l-1)}$	Number of (hidden) inputs fed into layer l
$M := h^{(0)}, C := h^{(L)}$	Input feature dimension, output dimension (number of classes for classification)
$\boldsymbol{z}^{(l-1)} \in \mathbb{R}^{h^{(l-1)}}$	(Hidden) features fed into layer l (output of layer $l-1$)
$\boldsymbol{x} := \boldsymbol{z}^{(0)}, \boldsymbol{f} := \boldsymbol{z}^{(L)}$	Input to the neural network, and its prediction for input $\boldsymbol{x}$
$f^{(l)}_{\boldsymbol{\theta}^{(l)}}$	Layer l parameterized by $\boldsymbol{\theta}^{(l)}$, mapping input $\boldsymbol{z}^{(l-1)}$ to output $\boldsymbol{z}^{(l)}$
$f_{\boldsymbol{\theta}} := f^{(L)}_{\boldsymbol{\theta}^{(L)}} \circ \ldots \circ f^{(1)}_{\boldsymbol{\theta}^{(1)}}$	Sequential feedforward neural network parameterized by $\boldsymbol{\theta}$, maps input $\boldsymbol{x}$ to output $\boldsymbol{f}$
$\boldsymbol{\theta}$	Parameter vector, concatenation of parameters over layers $\boldsymbol{\theta} := (\boldsymbol{\theta}^{(1)\top}, \ldots, \boldsymbol{\theta}^{(L)\top})^\top$

Derivatives

∇, J and ∇^2 denote the gradient, Jacobian, and Hessian, respectively.

$\mathrm{J}_{\boldsymbol{a}}\boldsymbol{b}$	Jacobian matrix of a vector $\boldsymbol{b}$ *w.r.t.* a vector $\boldsymbol{a}$, $[\mathrm{J}_{\boldsymbol{a}}\boldsymbol{b}]_{i,j} = \partial[\boldsymbol{b}]_i / \partial[\boldsymbol{a}]_j$
$\mathrm{J}_{\mathsf{A}}\mathsf{B}$	Generalized Jacobian matrix for tensor variables, $[\mathrm{J}_{\mathsf{A}}\mathsf{B}]_{i,j} = \partial[\operatorname{vec}\mathsf{B}]_i / \partial[\operatorname{vec}\mathsf{A}]_j$
$\nabla_{\boldsymbol{a}} b := (\mathrm{J}_{\boldsymbol{a}} b)^\top$	Gradient vector of a scalar b *w.r.t.* a vector $\boldsymbol{a}$, $[\nabla_{\boldsymbol{a}} b]_i = \partial b / \partial a_i$
$\nabla^2_{\boldsymbol{a}} b$	Hessian matrix of a scalar b *w.r.t.* a vector $\boldsymbol{a}$, $[\nabla^2_{\boldsymbol{a}} b]_{i,j} = \partial^2 b / \partial[\boldsymbol{a}]_i \partial[\boldsymbol{a}]_j$ (symmetric)
$\nabla^2_{\boldsymbol{a}}\boldsymbol{b}$, $\nabla^2_{\mathsf{A}}\mathsf{B}$	Generalized Hessian matrix (in general not quadratic, hence not symmetric) of a vector $\boldsymbol{b}$ *w.r.t.* a vector $\boldsymbol{a}$, or more general tensor variables
$\boldsymbol{g}_n(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta})$	Gradient of the loss implied by sample n
$\boldsymbol{g}_{\mathbb{D}}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbb{D}}(\boldsymbol{\theta})$	Gradient of the empirical risk implied by a dataset $\mathbb{D}$
$\boldsymbol{g}_{\mathbb{B}}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathbb{B}}(\boldsymbol{\theta})$	Mini-batch gradient
$\boldsymbol{H}_n(\boldsymbol{\theta}) := \nabla^2_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta})$	Hessian of the loss implied by sample n
$\boldsymbol{H}_{\mathbb{D}}(\boldsymbol{\theta}) := \nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathbb{D}}(\boldsymbol{\theta})$	Hessian of the empirical risk implied by a dataset $\mathbb{D}$
$\boldsymbol{H}_{\mathbb{B}}(\boldsymbol{\theta}) := \nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\mathbb{B}}(\boldsymbol{\theta})$	Mini-batch Hessian
$\boldsymbol{H}^{(l)}(\boldsymbol{\theta}^{(l)})$ or $\boldsymbol{H}(\boldsymbol{\theta}^{(l)})$	The block in the Hessian corresponding to layer l
$\boldsymbol{G}_{\mathbb{D}}(\boldsymbol{\theta})$	Generalized Gauss-Newton matrix on a dataset $\mathbb{D}$
$\boldsymbol{G}^{(l)}(\boldsymbol{\theta}^{(l)})$ or $\boldsymbol{G}(\boldsymbol{\theta}^{(l)})$	The block in the generalized Gauss-Newton matrix corresponding to layer l

Statistics

$\mathcal{U}(\{1, \ldots, N\})$	Uniform distribution over $\{1, \ldots, N\}$
$\mathcal{N}(x \mid \mu, \sigma^2)$	Uni-variate normal/Gaussian distribution of random variable x, with mean μ, positive variance σ^2, and density $\mathcal{N}(x \mid \mu, \sigma^2) = 1/\sigma\sqrt{2\pi} \exp[-1/2((x-\mu)/\sigma)^2]$
$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multi-variate normal/Gaussian distribution of random vector $\boldsymbol{x}$ with mean vector $\boldsymbol{\mu}$, PSD covariance matrix $\boldsymbol{\Sigma}$, and density $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = 1/(\sqrt{2\pi \det \boldsymbol{\Sigma}}) \exp[-1/2(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})]$
$\mathrm{Cat}(c \mid \boldsymbol{p})$	Multinomial/Categorical distribution with probabilities $\boldsymbol{p}$ for categories c

Miscellaneous

log	The natural logarithm (base e, *i.e.* $\log(\mathrm{e}) = 1$)

$\text{onehot}(c)$	One-hot vector of class c with $\text{onehot}(c) = \delta_{i,c}$
$\text{softmax}(\boldsymbol{a})$	Softmax probabilities of the logits $\boldsymbol{a}$, $[\text{softmax}(\boldsymbol{a})]_c = {}^{\exp(a_c)}/_{\sum_{i=1}\exp(a_i)}$.
$\delta_{i,j}$, $\delta(\boldsymbol{x} - \boldsymbol{a})$	Kronecker delta ($\delta_{i,i} = 1$ and $\delta_{i,j\neq i} = 0$), Dirac delta distribution
$(\mathbb{X} \to \mathbb{Y})$	Signature of a function that maps between $\mathbb{X}$ and $\mathbb{Y}$
$\{\boldsymbol{x}_n\}$ or $\{\boldsymbol{x}_n\}_n$	A set/collection of vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots$ over the index set implied by n
$\hat{\boldsymbol{e}}_i$	Unit vector in direction i, *i.e.* $\hat{\boldsymbol{e}}_i = \text{onehot}(i)$
$\mathbf{1}_m$	An m-dimensional vector containing ones everywhere
$\log(\boldsymbol{a}), \exp(\boldsymbol{a})$	Elementwise natural logarithm and exponential function of a vector
$m_{\boldsymbol{\theta}_t}(\boldsymbol{\theta})$	Local approximation of the loss in around $\boldsymbol{\theta}_t$

Acronyms & Abbreviations

E.g. or *e.g.*	For example (*exempli gratia*)
Etc. or *etc.*	And so on (*et cetera*)
I.e. or *i.e.*	That is (*id est*)
I.i.d. or *i.i.d.*	Independent and identically distributed
W.r.t. or *w.r.t.*	With respect to
AD	Automatic differentiation
API	Application Programming Interface
BDA	Block diagonal approximation
CG	Conjugate gradients
CNN	Convolutional neural network
CPU	Central processing unit
DNN	Deep neural network
DP	Differential privacy
FCNN	Fully-connected neural network
GGN	Generalized Gauss-Newton (matrix)
GN	Gauss-Newton (matrix)
GPU	Graphics processing unit
HBP	Hessian backpropagation
JMP	Jacobian-matrix product
JVP	Jacobian-vector product
KFAC	Kronecker-factored curvature
KFC	Kronecker factors for convolution
KFLR	Kronecker-factored low rank
KFRA	Kronecker-factored recursive approximation
MAP	Maximum a posteriori (estimation)
MC	Monte Carlo
MJP	Matrix-Jacobian product
ML	Machine learning
MLE	Maximum likelihood estimation
MLP	Multi-layer perceptron
NGD	Natural gradient descent
PCH	Positive-curvature Hessian
PD	Positive definite
PSD	Positive semi-definite
ResNet	Residual (neural) network
SNR	Signal-to-noise ratio
TPU	Tensor processing unit
VJP	Vector-Jacobian product

Overview 1.

1.1 Introduction

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift –

not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Procedure 1.1: **Canonical deep learning training loop.** After setting up the data, model, loss function, and optimizer, iterate over batches: in each iteration, compute the mini-batch loss in a forward pass, and its gradient with a backward pass. Then use the gradient as learning signal to update the model parameters.

```
1  % dataset = ...    # Learning task examples
2
3  model = ...       # Practitioner's choice
4  loss_func = ...  # Practitioner's choice
5
6  optimizer = ...  # First-order method
7
8  while not_converged: # Standard training loop
9  features, targets = dataset.next_minibatch()
10
11 # Forward pass: Compute the loss
12 predictions = model(features)
13 loss = loss_func(predictions, targets)
14
15 # Backward pass: Compute the gradient
16 loss.backward()
17
18 # Update model parameters using the gradient
19 optimizer.step()
20 optimizer.zero_grad()
```

Reference to Line 16.

1.2 Outline

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

You have to run one the script `build-force.sh` to generate the following externalized TikZ figure:

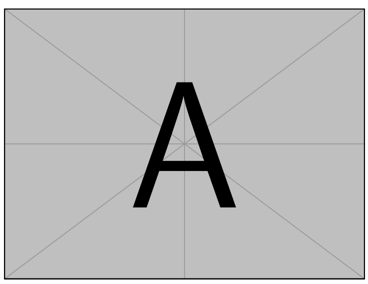
Figure 1.1: A TikZ figure: This figure checks if externalization works.

Disclaimer 1.1 Chapter 4 is based on the peer-reviewed conference publication with the following co-author contributions:

F. Dangel, F. Kunstner, and P. Hennig. "BackPACK: Packing more into Backprop". *International Conference on Learning Representations (ICLR)*. 2020 [1]

	Ideas	**Experiments**	**Analysis**	**Writing**
F. Dangel	33 %	55 %	45 %	35 %
F. Kunstner	33 %	45 %	45 %	45 %
P. Hennig	33 %	0 %	10 %	20 %

Chapter 4: BackPACK: an efficient framework built on top of PyTorch that extends the backpropagation algorithm.



github.com/f-dangel/backpack

Part I.

Background & Motivation

Background 1 2.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

The arrangement of partial derivatives in the generalizations of Jacobian and Hessian implies the following chain rule generalization for second-order derivatives:

Theorem 2.1 (Chain rule for the generalized Hessian) Let $\boldsymbol{b} : \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{c} : \mathbb{R}^m \to \mathbb{R}^p$ be twice differentiable and $\boldsymbol{d} = \boldsymbol{c} \circ \boldsymbol{b} : \mathbb{R}^n \to \mathbb{R}^p, \boldsymbol{a} \mapsto \boldsymbol{d}(\boldsymbol{a}) = \boldsymbol{c}(\boldsymbol{b}(\boldsymbol{a}))$. The relation between the Hessian of $\boldsymbol{d}$ and the Jacobians and Hessians of the constituents $\boldsymbol{c}$ and $\boldsymbol{b}$ is given by

$$\begin{aligned}&\nabla_{\boldsymbol{a}}^2 \boldsymbol{d}(\boldsymbol{a}) \\ &\quad = \left[\boldsymbol{I}_p \otimes \mathrm{J}_{\boldsymbol{a}} \boldsymbol{b}(\boldsymbol{a})\right]^\top \left[\nabla_{\boldsymbol{b}}^2 \boldsymbol{c}(\boldsymbol{b})\right] \mathrm{J}_{\boldsymbol{a}} \boldsymbol{b}(\boldsymbol{a}) \\ &\quad + \left[\mathrm{J}_{\boldsymbol{b}} \boldsymbol{c}(\boldsymbol{b}) \otimes \boldsymbol{I}_n\right] \nabla_{\boldsymbol{a}}^2 \boldsymbol{b}(\boldsymbol{a})\end{aligned} \tag{2.1}$$

2.1 Background topic 1

Definition 2.1 (Hessian) Let $b : \mathbb{R}^D \to \mathbb{R}; \boldsymbol{a} \mapsto b(\boldsymbol{a})$ be a differentiable vector-to-scalar function. The Hessian $\nabla_{\boldsymbol{a}}^2 b \in \mathbb{R}^{D \times D}$ of b *w.r.t.* $\boldsymbol{a}$ is a symmetric matrix containing the second-order partial derivatives

$$\nabla_{\boldsymbol{a}}^2 b = \frac{\partial^2 b}{\partial \boldsymbol{a} \partial \boldsymbol{a}^\top} \qquad \text{with} \qquad [\nabla_{\boldsymbol{a}}^2 b]_{i,j} = \frac{\partial^2 b}{\partial a_i \partial a_j} \tag{2.2}$$

The Hessian will often be denoted by $\boldsymbol{H}$. *E.g.* $\boldsymbol{H}_{p_{\text{data}}}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_{p_{\text{data}}}(\boldsymbol{\theta})$ for the Hessian of the population risk, and $\boldsymbol{H}_{\mathbb{D}}(\boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_{\mathbb{D}}(\boldsymbol{\theta})$ for the Hessian of the empirical risk on a dataset $\mathbb{D}$ (with $\mathbb{D} = \mathbb{D}_{\text{train}}, \mathbb{B}$ for the train loss and mini-batch Hessian).

2.2 Background topic 2

Example 2.1 (Least squares regression & square loss) Regression associates features in $\mathbb{X} = \mathbb{R}^M$ with targets in $\mathbb{Y} = \mathbb{R}^C$. A prediction in $\mathbb{F} = \mathbb{R}^C$ compares to its ground truth via the mean squared error[1]

$$\ell(\boldsymbol{f}, \boldsymbol{y}) = \frac{1}{C} \sum_{c=1}^{C} (y_c - f_c)^2 = \frac{1}{C} \|\boldsymbol{y} - \boldsymbol{f}\|_2^2 \tag{2.3}$$

1: There exist different conventions for the normalization factor. This text adapts the implementation of `MSELoss` (with `reduction="mean"` mode) in PyTorch for consistency with the code presented in later chapters. Normalizing by $^1/_C$ is also close to what the name, mean squared error, suggests.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Remark 2.1 (The log-probability's θ-gradient vanishes in expectation)

$$\begin{aligned}&-\int_\Omega p_\theta(z) \nabla_\theta \log p_\theta(z)\, \mathrm{d}z \\ &\qquad = -\int_\Omega p_\theta(z) \frac{\nabla_\theta p_\theta(z)}{p_\theta(z)}\, \mathrm{d}z \\ &\qquad = -\nabla_\theta \left(\int_\Omega p_\theta(z)\, \mathrm{d}z\right) \\ &\qquad = -\nabla_\theta 1 = 0\end{aligned}$$

Table 2.1: Forward pass for common modules used in feedforward networks. Input and output are denoted $\boldsymbol{x}, \boldsymbol{z}$ rather than $\boldsymbol{z}^{(l)}, \boldsymbol{z}^{(l+1)}$ to avoid clutter. $\boldsymbol{I}$ is the identity matrix. Bold upper-case symbols ($\boldsymbol{W}, \boldsymbol{X}, \boldsymbol{Z}, \dots$) denote matrices and bold upper-case sans serif symbols ($\mathsf{\mathbf{W}}, \mathsf{\mathbf{X}}, \mathsf{\mathbf{Z}}, \dots$) denote tensors.

OPERATION	FORWARD
Matrix-vector multiplication	$\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{W}) = \boldsymbol{W}\boldsymbol{x}$
Matrix-matrix multiplication	$\boldsymbol{Z}(\boldsymbol{X}, \boldsymbol{W}) = \boldsymbol{W}\boldsymbol{X}$
Addition	$\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{b}) = \boldsymbol{x} + \boldsymbol{b}$
Elementwise activation	$\boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})$, s.t. $z_i(\boldsymbol{x}) = \phi(x_i)$
Skip-connection	$\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x} + \boldsymbol{s}(\boldsymbol{x}, \boldsymbol{\theta})$
Reshape/view	$\mathsf{\mathbf{Z}}(\mathsf{\mathbf{X}}) = \operatorname{reshape}(\mathsf{\mathbf{X}})$
Index select/map π	$\boldsymbol{z}(\boldsymbol{x}) = \boldsymbol{\Pi}\boldsymbol{x}$, $\Pi_{j,\pi(j)} = 1$,
Convolution	$\mathsf{\mathbf{Z}}(\mathsf{\mathbf{X}}, \mathsf{\mathbf{W}}) = \mathsf{\mathbf{X}} \star \mathsf{\mathbf{W}}$,
	$\boldsymbol{Z}(\boldsymbol{W}, [\![\mathsf{\mathbf{X}}]\!]) = \boldsymbol{W}[\![\mathsf{\mathbf{X}}]\!]$,
Square loss	$\ell(\boldsymbol{f}, \boldsymbol{y}) = 1/C(\boldsymbol{y} - \boldsymbol{f})^\top(\boldsymbol{y} - \boldsymbol{f})$
Softmax cross-entropy	$\ell(\boldsymbol{f}, y) = -\operatorname{onehot}(y)^\top \log[\boldsymbol{p}(\boldsymbol{f})]$

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3. Background 2 (advanced)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

3.1 Background topic 1 (advanced)

3.2 Background topic 2 (advanced)

Part II.

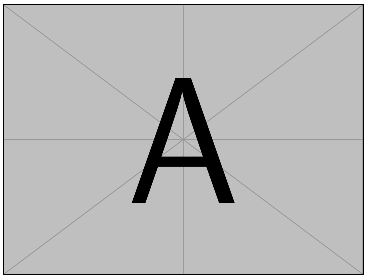
Backpropagation Beyond the Gradient

Paper 1 | 4.

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Code and experiments available at the Github repositories `f-dangel/backpack`, `f-dangel/backpack-experiments`



4.1 Introduction

4.2 Theory & Implementation

4.3 Evaluation & Benchmarks

4.4 Experiments

4.5 Conclusion

Paper 2 | 5.

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

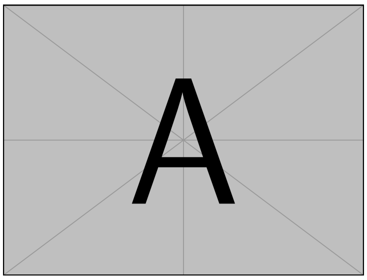
And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text

should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Code and experiments available at the Github repositories `f-dangel/cockpit`, `f-dangel/cockpit-experiments`



5.1 Introduction & Motivation

5.2 Cockpit's Instruments

5.3 Experiments

5.4 Showcase

5.5 Benchmark

5.6 Conclusion

Acknowledgments

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Part III.

Conclusion & Future Directions

6. Conclusion & Future Directions

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

6.1 Summary & Impact

Extending Backpropagation to the Hessian

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Packing More into Backprop

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Enabling a Closer Look Into Neural Nets

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Enabling Novel Ways to Compute with Curvature

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

6.2 Future Work

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Extending Cockpit

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Noise-aware Second-order Methods

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Optimizing Run Time & Advancing Automatic Differentiation

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Part IV.

Appendix

A. Additional Material for Chapter 4

B. Additional Material for Chapter 5

Bibliography

[1] F. Dangel, F. Kunstner, and P. Hennig. “BackPACK: Packing more into Backprop”. *International Conference on Learning Representations (ICLR)*. 2020.

[2] I. J. Goodfellow, Y. Bengio, and A. Courville. “Deep Learning”. 2016.