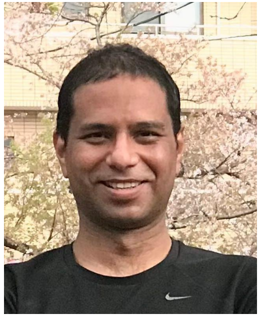
Duality Principles for Modern Machine Learning


Zelda Mariet


Mathieu Blondel


Thomas Möllenhoff


Emtiyaz Khan

Google DeepMind

Supported by the Bayes-Duality project,
JST CREST Grant Number JPMJCR2112

RIKEN AIP

Volunteers / Logistics

Peter Nickl
Research Assistant,
RIKEN AIP

Rob Brekelmans
Postdoctoral Fellow,
Vector Institute

Advisory Committee

Suvrit Sra
Professor
MIT

Francis Bach
Researcher
INRIA / ENS

Nihat Ay
Professor
TU Hamburg

Duality?

DUALITY IN MATHEMATICS AND PHYSICS*

SIR MICHAEL F. ATIYAH

ABSTRACT. Duality is one of the oldest and most fruitful ideas in Mathematics. I will survey its history, showing how it has constantly been generalized and has guided the development of Mathematics. I will bring it up to date by discussing some of the most recent ideas and conjectures in both Mathematics and Physics.

Michael F. Atiyah (1929-2019)

INTRODUCTORY REMARKS

Duality in mathematics is not a theorem, but a "principle". It has a simple origin, it is very powerful and useful, and has a long history going back hundreds of years. Over time it has been adapted and modified and so we can still use it in novel situations. It appears in many subjects in mathematics (geometry, algebra, analysis) and in physics. Fundamentally, duality gives *two different points of view of looking at the same object.* There are many things that have two different points of view and in principle they are all dualities.

Dual problems in convex optimization

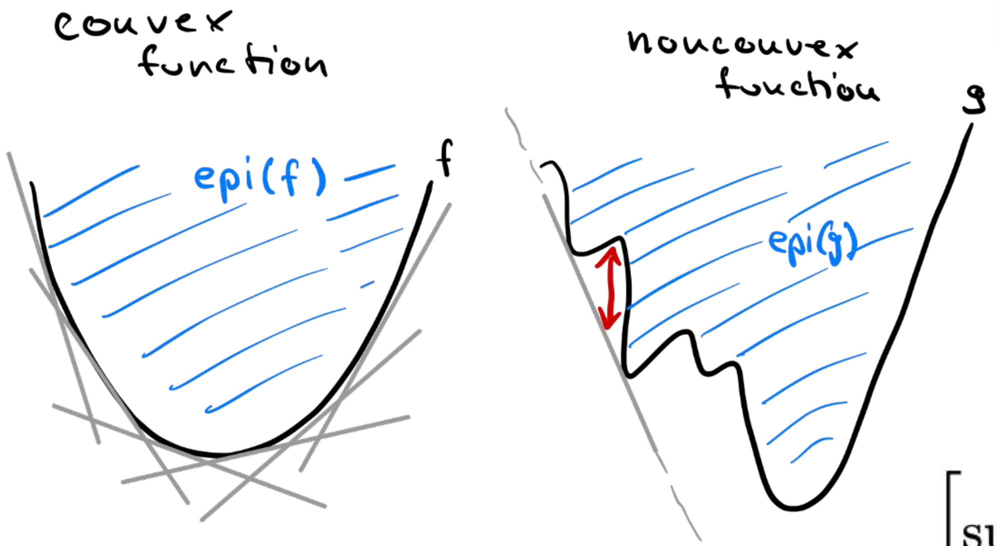
$$\text{(P)} \qquad \text{minimize} \quad f(x) - g(Ax) \quad \text{over} \quad x \in E\,,$$

$$\text{(P*)} \qquad \text{maximize} \quad g^*(y^*) - f^*(A^*y^*) \quad \text{over} \quad y^* \in F^*\,.$$

$$\begin{tikzcd}
E \arrow[r, "A"] & F \arrow[d, "\partial g"] \\
E^* \arrow[u, "\partial f^*"] & F^* \arrow[l, "A^*"]
\end{tikzcd}$$

[1] R. T. Rockafellar, Duality and stability in extremum problems involving convex function, Pacific J. Math. 1967

Duality for Convex Relaxation of Nonconvex Functions



$$\mathbb{E}_{\boldsymbol{\epsilon}'\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\left[\ell(\mathbf{m}+\boldsymbol{\epsilon}')\right] + \frac{\delta}{2}\|\mathbf{m}\|^2$$

$$\Big\downarrow \quad \textbf{Duality}$$

$$\left[\sup_{\boldsymbol{\epsilon}} \ell(\mathbf{m}+\boldsymbol{\epsilon}) - \frac{1}{2\sigma^2}\|\boldsymbol{\epsilon}\|^2\right] + \frac{\delta'}{2}\|\mathbf{m}\|^2$$

[1] Foret et al., Sharpness-Aware Minimization for Efficiently Improving Generalization, ICLR 2021.
[2] Thomas Möllenhoff, Mohammad Emtiyaz Khan. SAM as an Optimal Relaxation of Bayes, ICLR 2023.

I learned about duality back in 2008

Picture from 2009 in XRCE, Grenoble, France

Fast Dual Variational Inference for Non-Conjugate Latent Gaussian Models

Mohammad Emtiyaz Khan EMTIYAZ.KHAN@EPFL.CH
School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland

“Won’t work because it relies on convexity”

Michael P. Friedlander MPF@CS.UBC.CA
Department of Computer Science, University of British Columbia, Vancouver, Canada

Matthias Seeger MATTHIAS.SEEGER@EPFL.CH
School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne, Switzerland

Abstract

Latent Gaussian models (LGMs) are widely

1. Introduction

Latent Gaussian models (LGM) are ubiquitous in machine learning and statistics (e.g., Gaussian process

Duality connects many things!

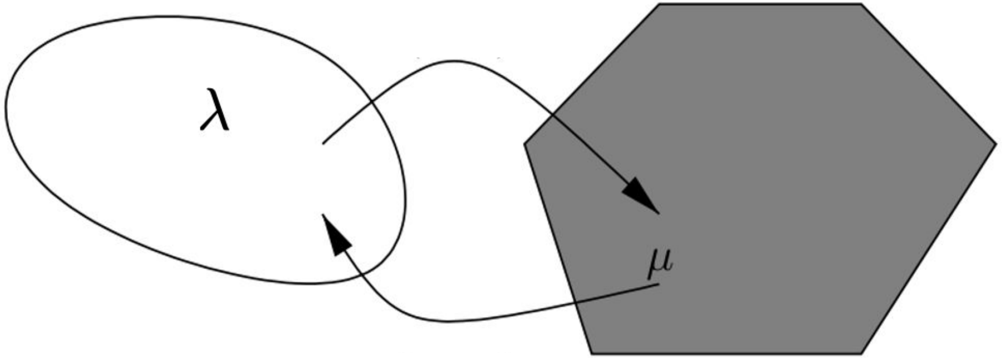
Bayesian Learning Rule [1]: Unify many algorithms in deep learning, optimization, and inference.

$$\lambda \leftarrow (1-\rho)\lambda - \rho \nabla_{\mu} \mathbb{E}_{q_\mu}[\text{loss}]$$

Key idea: the duality of Exponential Family [2]

Natural parameter

Expectation parameter



Natural Gradient Descent
(information geometry [4])

Mirror Descent [3]
(optimization)

Bayes' rule

Maximum-Entropy principle

1. Khan and Rue, Bayesian Learning rule, 2021
2. Wainwright and Jordan, Graphical models, exp-family, and variational inference, 2006
3. Raskutti and Mukherjee, Information Geometry of Mirror Descent, 2015
4. Amari, Information Geometry and its applications, 2016

How to represent and adapt the knowledge?

Sensitivity, perturbation, duality[1,2]

Based on the Bayesian Learning Rule, we are now developing a new notion of duality called the Bayes-Duality.

Check out these related features talks and posters

- Talks by Ehsan Amid, Len Spek, Ronny Bergmann
- Poster: The Memory-Perturbation Equation: Understanding Model's Sensitivity to Data
- Poster: Memory Maps to Understand Models
- Poster: Sparse Function-Space Representation of Neural Networks

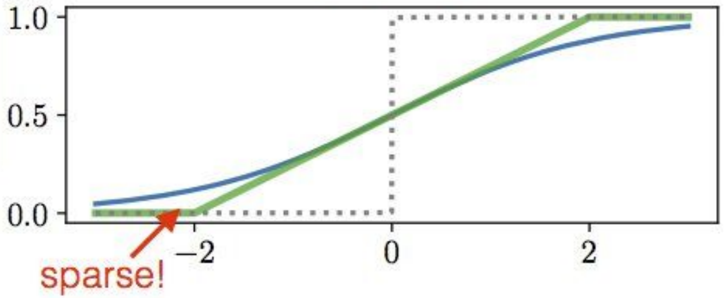
1. Schölkopf et al., A generalized representer theorem, 2001
2. Kimeldorf and Wahba, A correspondence between Bayes on stochastic process…, 1970

Prediction function regularized by Ω:

$$\hat{\mathbf{y}}_\Omega(\boldsymbol{\theta}) := \mathrm{argmax}_{\mathbf{p}\in\Delta} \ \langle \boldsymbol{\theta}, \mathbf{p} \rangle \ - \Omega(\mathbf{p})$$

Probability simplex: $\Delta := \{\mathbf{p} \in \mathbb{R}^d : \|\mathbf{p}\|_1 = 1, \mathbf{p} \geq 0\}$

$\hat{\mathbf{y}}_\Omega([s,0])_1$



Fenchel-Young loss generated by Ω:

$$L_\Omega(\boldsymbol{\theta}; \mathbf{y}) := \Omega^*(\boldsymbol{\theta}) + \Omega(\mathbf{y}) - \langle \boldsymbol{\theta}, \mathbf{y} \rangle$$

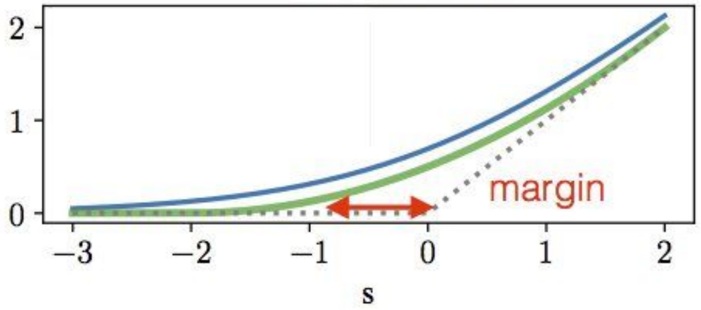
Fenchel conjugate of Ω restricted to Δ:

$\Omega^*(\boldsymbol{\theta}) := \max_{\mathbf{p}\in\Delta} \ \langle \boldsymbol{\theta}, \mathbf{p} \rangle \ - \Omega(\mathbf{p})$

$\mathbf{y}_{\text{neg}} := [0,1]^T$

$\mathbf{y}_{\text{pos}} := [1,0]^T$

$L_\Omega([s,0]; \mathbf{y}_{\text{neg}})$



Proposition:

$$L_\Omega(\boldsymbol{\theta}; \mathbf{y}) = 0 \Leftrightarrow \hat{y}_\Omega(\boldsymbol{\theta}) = \mathbf{y}$$

argmax: $\Omega = 0$
softmax: $\Omega(p) = \Sigma_j \, p_j \log p_j$
sparsemax: $\Omega(\mathbf{p}) = \|\mathbf{p}\|^2$

"Learning with Fenchel-Young losses". Blondel, Martins, Niculae. JMLR, 2020.

Duality as a natural space for updates and manipulations

$$\mathbb{E}\left[(y - f(x))^2\right] = \underbrace{\mathbb{E}\left[(y - \mathbb{E}y)^2\right]}_{\text{label noise}} + \underbrace{\left(\mathbb{E}y - \mathbb{E}f(x)\right)^2}_{\text{bias}} + \underbrace{\mathbb{E}\left[\left(f(x) - \mathbb{E}f(x)\right)^2\right]}_{\text{variance}}$$

Several generalization attempts, but none as elegant as the MSE - unless we turn to duality!

Bregman divergence losses introduce a dual space which naturally surfaces the decomposition

$$\mathbb{E}\Big[D[y \,\|\, f(x)]\Big] = \mathbb{E}\Big[D[y \,\|\, \mathbb{E}y]\Big] + D\Big[\mathbb{E}y \,\|\, \left(\mathbb{E}f(x)^*\right)^*\Big] + \mathbb{E}\Big[D[\left(\mathbb{E}f(x)^*\right)^* \| \, f(x)]\Big]$$

Tl;dr: Natural space for labels = **primal**; natural space for predictions = **dual**.

Pfau, 2013; Gupta et al, 2022

Time	Event	Title	Speaker
9:00 - 9:30	Opening Remarks	Duality: Opening Remarks	Workshop Organizers
9:30 - 9:55	Invited Talk	Fenchel Duality Theory on Riemannian Manifolds and the Riemannian Chambolle-Pock Algorithm	Ronny Bergmann
9:55 - 10:05	**Coffee break**		
10:05 - 10:17	Contributed Talk	Time-Reversed Dissipation Induces Duality Between Minimizing Gradient Norm and Function Value	Jaeyeon Kim
10:17 - 10:29	Contributed Talk	RIFLE: Imputation and Robust Inference from Low Order Marginals	Sina Baharlouei
10:30 - 10:42	Contributed Talk	A Representer Theorem for Vector-Valued Neural Networks: Insights on Weight Decay Training and Widths of Deep Neural Networks	Joseph Shenouda
10:45 - 11:10	Invited Talk	Duality from Gradient Flow Force-Balance to Distributionally Robust Learning	Jia-Jie Zhu
11:10 - 11:35	Invited Talk	Convergence of mean field Langevin dynamics: Duality viewpoint and neural network optimization	Taiji Suzuki

11:35 - 12:00	Invited Talk	Duality for Neural Networks through Reproducing Kernel Banach Spaces	Len Spek
12:00 - 13:00	**Lunch**		
13:00 - 14:30	**Poster Session**		
14:30 - 14:55	Invited Talk	Dual RL: Unification and New Methods for Reinforcement and Imitation Learning	Amy Zhang
14:55 - 15:35	**Coffee Break & Poster Session**		
15:35 - 16:00	Invited Talk	A Dualistic View of Activations in Deep Neural Networks	Ehsan Amid
16:00 - 17:00	Panel Discussion	Duality in Modern Machine Learning	Speakers & Organizers

Duality!