RISK-AVERSE DISTRIBUTIONAL REINFORCEMENT LEARNING

A CVAR OPTIMIZATION APPROACH

SILVESTR STANKO

Department of Computer Science
Faculty of Electrical Engineering
Czech Technical University

January, 2018

ABSTRACT

Short summary of the contents in English. . . a great guide by Kent Beck how to write good abstracts can be found here:

`https://plg.uwaterloo.ca/~migod/research/beckOOPSLA.html`

ABSTRAKT

Český abstrakt

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— **knuth:1974** [**knuth:1974**]

ACKNOWLEDGMENTS

Put your acknowledgments here.

Many thanks to everybody who already sent me a postcard!

Regarding the typography and other help, many thanks go to Marco Kuhlmann, Philipp Lehman, Lothar Schlesier, Jim Young, Lorenzo Pantieri and Enrico Gregorio[1], Jörg Sommer, Joachim Köstler, Daniel Gottschlag, Denis Aydin, Paride Legovini, Steffen Prochnow, Nicolas Repp, Hinrich Harms, Roland Winkler, Jörg Weber, Henri Menke, Claus Lahiri, Clemens Niederberger, Stefano Bragaglia, Jörn Hees, Scott Lowe, Dave Howcroft, and the whole LaTeX-community for support, ideas and some great software.

Regarding LyX: The LyX port was intially done by *Nicholas Mariette* in March 2009 and continued by *Ivo Pletikosić* in 2011. Thank you very much for your work and for the contributions to the original style.

1 Members of GuIT (Gruppo Italiano Utilizzatori di TeX e LaTeX)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

1

INTRODUCTION

We consider the problem of maximizing some notion of reward in a Markov decision process (MDP). Contrary to the usual case of maximizing the expected discounted reward, we focus on maximizing a risk-sensitive objective, which takes into account the variability of the reward and allows us to avoid catastrophic events. Risk-sensitivity has also recently been shown to capture the robustness to modeling errors [7] [more], providing us with further motivation. Risk-sensitivity can be modeled by replacing the risk-neutral expectation by an alternate risk-measure of the total discounted reward. In this work we consider risk-sensitive MDPs with a Conditional Value-at-Risk (CVaR) objective, a risk-measure which has been recently gaining popularity [cite] due to it's favorable computational properties [cite].

To do this, we utilize recent distributional Reinforcement Learning (RL) advances [4][8], that replace the usual Q-function by a distribution of the discounted reward, allowing us to gain more information about the structure of the return, which can be than utilized in maximizing an alternative objective.

1.1 MOTIVATION

1.1.1 *xxx*

1.2 CONTRIBUTIONS

1.3 THESIS OUTLINE

2

PRELIMINARIES

Let Z be a bounded-mean random variable, i.e. $\mathbb{E}[|Z|] < \infty$, with cummulative distribution function (c.d.f.) $F(z) = \mathbb{P}(Z \leq z)$. In this paper we interpret Z as a reward[1]. The value-at-risk (VaR) at confidence level $\alpha \in (0, 1)$ is the α quantile of Z, i.e.

$$\mathrm{VaR}_\alpha(Z) = F^{-1}(\alpha) = \max\{z | F(z) \leq \alpha\} \tag{2.1}$$

We will use the notation $\mathrm{VaR}_\alpha(Z)$, $F^{-1}(\alpha)$ interchangebly, often explictly denoting the random variable of inverse c.d.f. as $F_Z^{-1}(\alpha)$.

2.0.1 *Conditional Value-at-Risk*

The conditional value-at-risk (CVaR) at confidence level $\alpha \in (0, 1)$ is defined as:

$$\mathrm{CVaR}_\alpha(Z) = \frac{1}{\alpha}\int_0^\alpha F_Z^{-1}(\beta)\mathrm{d}\beta = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_\beta(Z)\mathrm{d}\beta \tag{2.2}$$

We will also use the following equivalent formulation from [17]:

$$\mathrm{CVaR}_\alpha(Z) = \max_s \left\{ \frac{1}{\alpha}\mathbb{E}\left[(Z - s)^-\right] + s \right\} \tag{2.3}$$

where $(x)^- = \min(x, 0)$ represents the negative part of x.

2.0.2 *Markov Decision Processes*

An MDP is a 5-tuple $\mathcal{M} = (\mathcal{X}, \mathcal{A}, R, P, \gamma)$, where $\mathcal{X}$ and $\mathcal{A}$ are the finite state and action spaces; $R(x, a) \in [R_{\min}, R_{\max}]$ is a random variable representing the reward generated by being in state x and selecting action a; $P(\cdot|x, a)$ is the transition probability distribution; $\gamma \in [0, 1)$ is a discount factor. We also assume we are given a starting state x_0.

A stationary (or markovian) policy is a mapping from states to actions $\pi : \mathcal{X} \to \mathcal{A}$.

1 This is in accordance with reinforcement learning literature and opposed to risk-related literature.

We define the return $Z^\pi(x_t)$ as a random variable representing the discounted reward along a trajectory generated by the MDP by following the policy π, starting at state x_t

$$Z^\pi(x_t) = \sum_{t=0}^{\infty} \gamma^t R(x_t, \pi(x_t)) \tag{2.4}$$

We will sometimes omit the π superscript when the policy is clear from context.

2.0.3 *Bellman equation*

The Bellman equation is a recursive equation that defines the action-value function Q:

$$Q^\pi(x,a) = \mathbb{E}\left[Z^\pi(x,a)\right] = \mathbb{E}\left[Z^\pi(x,a)\right] \tag{2.5}$$

where the next state X' is sampled according to the MDP's transition probabilities.
distributional

2.0.4 *Problem formulation*

The risk-sensitive problem we wish to adress for a given confidence level α is as follows:

$$\max_\pi \mathrm{CVaR}_\alpha(Z^\pi(x_0)) \tag{2.6}$$

2.0.5 *Time-consistency*

An important property of CVaR MDPs is that of time-consistency. ***Choose one definition, describe*** The notion of time consistency varies from author to author; in Shapiro [XXX] it is shown ***.

*** Policy gradient literature ignores the time consistency-issue, leading to locally optimal policies *** show that they can be worse than EXP ***.

2.1 REINFORCEMENT LEARNING

2.1.1 *rl*

2.1.2 *mdp*

2.1.3 *pi, vi, q-learning (?)*

2.1.4 *Distributional RL*

2.2 RISK-AVERSION

2.2.1 *general*

2.2.2 *var*

2.2.3 *cvar*

2.3 LITERATURE SURVEY

3

VALUE ITERATION WITH CVAR

Chow et al. [7] have recently proposed an approximate value iteration algorithm for CVaR MDPs. This algorithm requires the computation of a linear program in each step of the value iteration procedure. We utilize a connection between the used αCVaR_α function and the quantile function and propose a linear-time algorithm that substitutes the LP computation, making the CVaR value iteration feasible for much larger MDPs. ***experiments*** The procedure also opens the door for a Q-learning variant of the algorithm.

3.1 CVAR BELLMAN EQUATION

3.2 EFFICIENT COMPUTATION

3.2.1 *CVaR Value Iteration*

The results of [7] heavily rely on the CVaR decomposition theorem, which we show below

$$\begin{aligned} CVaR_\alpha(x,a) = \min_\xi & \sum_{x'} p(x,a,x')\xi(x')CVaR_{\xi\alpha}(x') \\ \text{s.t.} \quad & \sum_{x'} p(x,a,x')\xi(x') = 1 \\ & 0 \le \xi(x') \le \frac{1}{\alpha} \end{aligned} \tag{3.1}$$

The theorem states that we can compute the $CVaR_\alpha(x)$ as the minimal weighted combination of $CVaR_\alpha(x')$ under a perturbed distribution. [7] extended this result to a dynamic programming formulation

$$CVaR_\alpha(x) = \max_a R(x,a) + \gamma CVaR_\alpha(x,a) \tag{3.2}$$

and showed that the VI procedure is a contraction and perserves the convexity of $\alpha CVaR_\alpha$. The fixed point of this contraction is then the exact CVaR value.

This algorithm is unfortunately unusable in practice, as the state-space is continuous in α. The solution proposed in [7] is then to represent the convex $\alpha CVaR_\alpha$ as a piecwise linear function.

I definition

The interpolated Bellman operator is then also a contraction and has a bounded error

$$
\begin{aligned}
CVaR_\alpha(x,a) = \min_\xi & \sum_{x'} p(x,a,x') \frac{I_{x'}(\alpha\xi(x'))}{\alpha} \\
\text{s.t.} \quad & \sum_{x'} p(x,a,x')\xi(x') = 1 \\
& 0 \le \xi(x') \le \frac{1}{\alpha}
\end{aligned}
\tag{3.3}
$$

3.2.2 *Quantile representation*

We use the following two facts: firstly, any disrete distribution function has a piece-wise linear $\alpha CVaR_\alpha$ function [17]; secondly the $\alpha CVaR_\alpha$ and the quantile function can be computed from each other by utilizing the relation

$$
\frac{\partial}{\partial\alpha} \alpha CVaR_\alpha(Z) = \frac{\partial}{\partial\alpha} \int_0^\alpha VaR_\beta(Z) d\beta = VaR_\alpha(Z) \tag{3.4}
$$

integration constant

We propose the following improvement: instead of using linear programming for the CVaR computation, we instead use the distributions represented by the $\alpha CVaR_\alpha$ function.

The computation of CVaR of a discrete probability mixture is a linear-time process as we show bellow. The general steps of the computation are as follows

1. transform $\alpha CVaR_\alpha$ of each possible state transition to a discrete probability distribution function
2. combine these to to a distribution representing the full state-action distribution
3. compute $\alpha CVaR_\alpha$ for all atoms

3.2.3 *Proof*

3.3 VAR-BASED POLICY IMPROVEMENT

3.3.1 *Policy Improvement*

Recall the primal definition of CVaR (***). Our goal can then be rewritten as

$$
\max_\pi CVaR_\alpha^\pi(Z) = \max_\pi \max_s \frac{1}{\alpha} \mathbb{E}\left[(Z^\pi - s)^-\right] + s \tag{3.5}
$$

It also holds [XXX] that for the maximum, it holds $s^* = \mathrm{VaR}_\alpha$

$$
\mathrm{CVaR}_\alpha(Z) = \max_s \left\{ \frac{1}{\alpha} \mathbb{E}\left[(Z-s)^-\right] + s \right\} = \frac{1}{\alpha} \mathbb{E}\left[(Z - \mathrm{VaR}_\alpha(Z))^-\right] + \mathrm{VaR}_\alpha(Z)
$$

$$ \tag{3.6} $$

The main idea of the algorithm 1, partially explored in [3], is as follows: If we knew the value s^* in the solution to equation (***), we could simplify the problem to maximize only

$$\max_\pi CVaR_\alpha(Z) = \max_\pi \frac{1}{\alpha}\mathbb{E}\left[(Z^\pi - s^*)^-\right] + s^* \tag{3.7}$$

Given that we have access to the return distributions, we can improve the policy by simply choosing an action that maximizes CVaR in the first state $a_0 = \arg\max_\pi \mathrm{CVaR}_\alpha(Z^\pi(x_0))$. We can then, as an approximation, set $s = VaR_\alpha(Z(x_0))$ and then only maximize the simpler criterion

$$\max_\pi CVaR_\alpha(Z) = \max_\pi \frac{1}{\alpha}\mathbb{E}\left[(Z^\pi - s^*)^-\right] + s^* \tag{3.8}$$

The algorithm can be seen as coordinate ascent; in the first phase (when we compute CVaR) we maximize w.r.t. s while keeping π fixed; in the second phase we fix s and maximize w.r.t. π.

In the following theorem, we show that this indeed leads to a monotonic improvement over the previous policy.

Theorem 1. *Let π be a fixed policy, $\alpha \in (0, 1]$. By following policy π' from algorithm 1, we will improve $CVaR_\alpha(Z)$ in expectation:*

$$CVaR_\alpha(Z^\pi) \leq CVaR_\alpha(Z^{\pi'})$$

Proof. Let s^* be a solution to eq 2.3. Then by optimizing $\max_\pi \frac{1}{\alpha}\mathbb{E}\left[(Z - s^*)^-\right]$, we will monotonely improve the optimization criterion 3.7.

$$CVaR_\alpha(Z^\pi) = \frac{1}{\alpha}\mathbb{E}\left[(Z^\pi - s^*)^-\right] + s^* \leq \max_{\pi'} \frac{1}{\alpha}\mathbb{E}\left[(Z^{\pi'} - s^*)^-\right] + s^*$$

Note the following facts:

$$Z_t = R_t + \gamma Z_{t+1} \tag{3.9}$$

$$\mathbb{E}\left[(Z_t - s)^-\right] = \mathbb{E}\left[(Z_t - s)\mathbb{1}(Z_t \leq s)\right] \tag{3.10}$$

$$\mathbb{E}[H(Z)] = \sum_i p_i \mathbb{E}[H(Z_i)] \tag{3.11}$$

The last equation holds if $Z \sim p_i$ is a probability mixture for any function H.

We can rewrite the criterion as

$$\begin{aligned}
\mathbb{E}\left[(Z_t - s)^-\right] = \mathbb{E}\left[(Z_t - s)\mathbb{1}(Z_t \le s)\right] &= \mathbb{E}\left[(R_t + \gamma Z_{t+1} - s)\mathbb{1}(Z_{t+1} \le \frac{s - R_t}{\gamma})\right] \\
&= \sum_{x_{t+1}, r_t} P(x_{t+1}, r_t \mid x_t, a)\mathbb{E}\left[(r_t + \gamma Z(x_{t+1}) - s)\mathbb{1}(Z(x_{t+1}) \le \frac{s - r_t}{\gamma})\right] \\
&= \sum_{x_{t+1}, r_t} P(x_{t+1}, r_t \mid x_t, a)\mathbb{E}\left[\gamma\left(Z(x_{t+1}) - \frac{s - r_t}{\gamma}\right)\mathbb{1}(Z(x_{t+1}) \le \frac{s - r_t}{\gamma})\right] \\
&= \gamma \sum_{x_{t+1}, r_t} P(x_{t+1}, r_t \mid x_t, a)\mathbb{E}\left[\left(Z(x_{t+1}) - \frac{s - r_t}{\gamma}\right)\mathbb{1}(Z(x_{t+1}) \le \frac{s - r_t}{\gamma})\right] \\
&= \gamma \sum_{x_{t+1}, r_t} P(x_{t+1}, r_t \mid x_t, a)\mathbb{E}\left[\left(Z(x_{t+1}) - \frac{s - r_t}{\gamma}\right)^-\right]
\end{aligned} \tag{3.12}$$

Now let's say we sampled reward $\hat{r}_t$ and state $\hat{x}_{t+1}$, we are still trying to find a policy π^* that maximizes

$$\begin{aligned}
\pi^* &= \arg\max_\pi \mathbb{E}\left[(Z_t - s)^- \mid \hat{x}_{t+1}, \hat{r}\right] \\
&= \arg\max_\pi \mathbb{E}\left[\left(Z(\hat{x}_{t+1}) - \frac{s - \hat{r}_t}{\gamma}\right)^-\right]
\end{aligned} \tag{3.13}$$

Where we ignored the unsampled states (since these are not a function of $\hat{x}_{t+1}$) and the multiplicative constant γ that will not affect the maximum argument.

At the starting state, we set $s = s^*$. At each following state we select an action according to equation 3.13. By induction we maximize the criterion **??** in each step.

□

Algorithm 1 VaR-based policy improvement

input α, x_0, γ
$a = \arg\max_a CVaR_\alpha(Z(x_0, a))$
$s = VaR_\alpha(Z(x_0, a))$
$x_t, r_t = \text{envTransition}(x_0, a)$
while x_t is not terminal **do**
  $s = \dfrac{s - r_t}{\gamma}$
  $a = \arg\max_a \mathbb{E}\left[(Z(x_t, a) - s)^-\right]$
  $x_t, r_t = \text{envTransition}(x_t, a)$
end while

Note that the resulting policy is nonstationary, however we do not need an extended state-space to follow this policy, it is only necessary to remember our previous s.

3.3.2 *Repeated policy improvement*

This policy then could be evaluated again by the distributional Q-learning procedures, however we

3.4 EXPERIMENTS

3.4.1 *Cliffworld*

3.4.2 *Atari?*

3.5 SUMMARY

3.6 ξ-COMPUTATION

CVaR decomposition formulation (based on the dual):

$$\begin{aligned} CVaR_\alpha(x,a) = \min_\xi & \sum_{x'} p(x,a,x')\xi(x')CVaR_{\xi\alpha}(x') \\ \text{s.t.} \quad & \sum_{x'} p(x,a,x')\xi(x') = 1 \\ & 0 \le \xi(x') \le \frac{1}{\alpha} \end{aligned} \tag{3.14}$$

Theorem 2. *Solution to minimization problem 3.14 can be computed without optimization by setting*

$$\xi(x') = \frac{F_{x'}(F_x^{-1}(\alpha))}{\alpha} \tag{3.15}$$

Proof. For simplification, we work only with two states: x' the actual sampled state and $\bar{x}'$ representing the other states. The equation then simplifies to

$$\begin{aligned} CVaR_\alpha(x,a) &= \min_\xi \; p\xi CVaR_{\xi\alpha}(x') + (1-p)\frac{1-p\xi}{1-p}CVaR_{\frac{1-p\xi}{1-p}\alpha}(\bar{x}') \\ &= \min_\xi \; p\xi CVaR_{\xi\alpha}(x') + (1-p\xi)CVaR_{\frac{1-p\xi}{1-p}\alpha}(\bar{x}') \end{aligned} \tag{3.16}$$

To find the min we first find the first derivative[1] w.r.t. ξ

$$\begin{aligned} \frac{\partial CVaR_\alpha}{\partial \xi} &= pCVaR_{\xi\alpha} + p\xi\frac{\partial CVaR_{\alpha\xi}}{\partial\xi} - pCVaR_{\frac{1-p\xi}{1-p}\alpha} + (1-p\xi)\frac{\partial CVaR_{\frac{1-p\xi}{1-p}\alpha}}{\partial\xi} \\ &= pCVaR_{\xi\alpha} + p\xi\left[\frac{1}{\xi}VaR_{\xi\alpha} - \frac{1}{\xi}CVaR_{\xi\alpha}\right] - pCVaR_{\frac{1-p\xi}{1-p}\alpha} + (1-p\xi)\left[\frac{p}{1-p\xi}CVaR_{\frac{1-p\xi}{1-p}\alpha} - \right. \\ &= pCVaR_{\xi\alpha} + pVaR_{\xi\alpha} - pCVaR_{\xi\alpha} - pCVaR_{\xi\alpha} - pCVaR_{\frac{1-p\xi}{1-p}\alpha} + CVaR_{\frac{1-p\xi}{1-p}\alpha} - pVaR_{\frac{1-p\xi}{1-p}\alpha} \\ &= pVaR_{\xi\alpha} - pVaR_{\frac{1-p\xi}{1-p}\alpha} \end{aligned} \tag{3.17}$$

1 We used the following facts:

$$\frac{\partial CVaR_{\alpha\xi}}{\partial\xi} = \frac{1}{\xi}VaR_{\xi\alpha} - \frac{1}{\xi}CVaR_{\xi\alpha} \qquad \frac{\partial CVaR_{\frac{1-p\xi}{1-p}\alpha}}{\partial\xi} = \frac{p}{1-p\xi}CVaR_{\frac{1-p\xi}{1-p}\alpha} - \frac{p}{1-p\xi}VaR_{\frac{1-p\xi}{1-p}\alpha}$$

By setting the derivative to 0 (to find the min), we get

$$VaR_{\xi\alpha}(x') = VaR_{\frac{1-p\xi}{1-p}\alpha}(\bar{x}') \tag{3.18}$$

By inserting claim 3.15 into 3.18 we get the symmetrical claim

$$\frac{1-p\xi}{1-p} = \xi(\bar{x}') = \frac{F_{\bar{x}'}(F_x^{-1}(\alpha))}{\alpha} \tag{3.19}$$

We rewrite 3.16 as (assuming ξ is the minimum point)

$$\begin{aligned}\frac{1}{\alpha}\int_0^{\alpha} F_x^{-1}(t)dt &= p\xi\frac{1}{\xi\alpha}\int_0^{\xi\alpha} F_{x'}^{-1}(t)dt + (1-p\xi)\frac{1-p}{(1-p\xi)\alpha}\int_0^{\frac{1-p\xi}{1-p}\alpha} F_{\bar{x}'}^{-1}(t) \\ &= p\frac{1}{\alpha}\int_0^{\xi\alpha} F_{x'}^{-1}(t)dt + (1-p)\frac{1}{\alpha}\int_0^{\frac{1-p\xi}{1-p}\alpha} F_{\bar{x}'}^{-1}(t)\end{aligned} \tag{3.20}$$

This must also hold if we multiply both sides by α

$$\int_0^{\alpha} F_x^{-1}(t)dt = p\int_0^{\xi\alpha} F_{x'}^{-1}(t)dt + (1-p)\int_0^{\frac{1-p\xi}{1-p}\alpha} F_{\bar{x}'}^{-1}(t) \tag{3.21}$$

And we take derivations w.r.t. α of both sides

$$F_x^{-1}(\alpha) = p\xi F_{x'}^{-1}(\xi\alpha) + (1-p\xi)F_{\bar{x}'}^{-1}(\frac{1-p\xi}{1-p}\alpha) \tag{3.22}$$

By inserting 3.15 we get

$$\begin{aligned}p\xi F_{x'}^{-1}(\xi\alpha) + (1-p)\xi_2 F_{\bar{x}'}^{-1}(\xi_2\alpha) &= p\xi F_{x'}^{-1}(F_{x'}(F_x^{-1}(\alpha))) + (1-p\xi)F_{\bar{x}'}^{-1}\left(F_{\bar{x}'}(F_x^{-1}(\alpha))\right) \\ &= p\xi F_x^{-1}(\alpha) + (1-p\xi)F_x^{-1}(\alpha) = F_x^{-1}(\alpha)\end{aligned} \tag{3.23}$$

We've shown that the proposed solution 3.15 satisfies the minimization constraint 3.18 (= is a minimal point) and satisfies the dual decomposition 3.14. (This has been shown only in the differentiated form)

□

4

Q-LEARNING WITH CVAR

4.1 TODO

4.2 TODO

4.3 TODO

4.4 EXPERIMENTS

5

APPROXIMATE Q-LEARNING WITH CVAR

5.1 TODO

5.2 TODO

5.3 TODO

5.4 EXPERIMENTS

6

CONCLUSION

Bäuerle and Ott [3] Bellemare, Dabney, and Munos [4] Chow et al. [7] Dabney et al. [8] Garcıa and Fernández [9] Majumdar and Pavone [13] Morimura et al. [14] Morimura et al. [15] Pflug and Pichler [16] Rockafellar and Uryasev [17] Rockafellar and Uryasev [18] Majumdar and Pavone [13] Leike et al. [12] Amodei et al. [1] Shapiro [19] Artzner et al. [2] Tamar et al. [21] Sutton and Barto [20] Watkins and Dayan [23] Bellman [5] Tsitsiklis [22] Boyd and Vandenberghe [6] Kreyszig [11] Koenker and Hallock [10]

A

INTRODUCTION

This bundle for LaTeX has two goals:

1. Provide students with an easy-to-use template for their Master's or PhD thesis. (Though it might also be used by other types of authors for reports, books, etc.)

2. Provide a classic, high-quality typographic style that is inspired by **bringhurst:2002**'s "*The Elements of Typographic Style*" [**bringhurst:2002**].

Risk-Averse Distributional Reinforcement Learning version 0.0

The bundle is configured to run with a *full* MiKTeX or TeXLive[1] installation right away and, therefore, it uses only freely available fonts. (Minion fans can easily adjust the style to their needs.)

People interested only in the nice style and not the whole bundle can now use the style stand-alone via the file `classicthesis.sty`. This works now also with "plain" LaTeX.

As of version 3.0, `classicthesis` can also be easily used with LyX[2] thanks to Nicholas Mariette and Ivo Pletikosić. The LyX version of this manual will contain more information on the details.

This should enable anyone with a basic knowledge of LaTeX 2_ε or LyX to produce beautiful documents without too much effort. In the end, this is my overall goal: more beautiful documents, especially theses, as I am tired of seeing so many ugly ones.

The whole template and the used style is released under the GNU General Public License.

If you like the style then I would appreciate a postcard:

André Miede
Detmolder Straße 32
31737 Rinteln
Germany

The postcards I received so far are available at:

http://postcards.miede.de

A well-balanced line width improves the legibility of the text. That's what typography is all about, right?

So far, many theses, some books, and several other publications have been typeset successfully with it. If you are interested in some typographic details behind it, enjoy Robert Bringhurst's wonderful book.

1 See the file `LISTOFFILES` for needed packages. Furthermore, `classicthesis` works with most other distributions and, thus, with most systems LaTeX is available for.
2 http://www.lyx.org

IMPORTANT NOTE: Some things of this style might look unusual at first glance, many people feel so in the beginning. However, all things are intentionally designed to be as they are, especially these:

- No bold fonts are used. Italics or spaced small caps do the job quite well.
- The size of the text body is intentionally shaped like it is. It supports both legibility and allows a reasonable amount of information to be on a page. And, no: the lines are not too short.
- The tables intentionally do not use vertical or double rules. See the documentation for the `booktabs` package for a nice discussion of this topic.[3]
- And last but not least, to provide the reader with a way easier access to page numbers in the table of contents, the page numbers are right behind the titles. Yes, they are *not* neatly aligned at the right side and they are *not* connected with dots that help the eye to bridge a distance that is not necessary. If you are still not convinced: is your reader interested in the page number or does she want to sum the numbers up?

Therefore, please do not break the beauty of the style by changing these things unless you really know what you are doing! Please.

YET ANOTHER IMPORTANT NOTE: Since `classicthesis`' first release in 2006, many things have changed in the LaTeX world. Trying to keep up-to-date, `classicthesis` grew and evolved into many directions, trying to stay (some kind of) stable and be compatible with its port to LyX. However, there are still many remains from older times in the code, many dirty workarounds here and there, and several other things I am absolutely not proud of (for example my unwise combination of KOMA and `titlesec` etc.).

An outlook into the future of `classicthesis`*.*

Currently, I am looking into how to completely re-design and re-implement `classicthesis` making it easier to maintain and to use. As a general idea, `classicthesis.sty` should be developed and distributed separately from the template bundle itself. Excellent spin-offs such as `arsclassica` could also be integrated (with permission by their authors) as format configurations. Also, current trends of `microtype`, `fontspec`, etc. should be included as well. As I am not really into deep LaTeX programming, I will reach out to the LaTeX community for their expertise and help.

A.1 ORGANIZATION

A very important factor for successful thesis writing is the organization of the material. This template suggests a structure as the following:

You can use these margins for summaries of the text body. . .

- `Chapters/` is where all the "real" content goes in separate files such as `Chapter01.tex` etc.
- `FrontBackMatter/` is where all the stuff goes that surrounds the "real" content, such as the acknowledgments, dedication, etc.

3 To be found online at http://mirror.ctan.org/macros/latex/contrib/booktabs/.

- `gfx/` is where you put all the graphics you use in the thesis. Maybe they should be organized into subfolders depending on the chapter they are used in, if you have a lot of graphics.
- `Bibliography.bib`: the BibTEX database to organize all the references you might want to cite.
- `classicthesis.sty`: the style definition to get this awesome look and feel. Does not only work with this thesis template but also on its own (see folder `Examples`). Bonus: works with both LATEX and PDFLATEX. . . and LYX. Great tool and it's free!
- `ClassicThesis.tex`: the main file of your thesis where all gets bundled together.
- `classicthesis-config.tex`: a central place to load all nifty packages that are used.

  Make your changes and adjustments here. This means that you specify here the options you want to load `classicthesis.sty` with. You also adjust the title of your thesis, your name, and all similar information here. Refer to Section A.3 for more information.

  This had to change as of version 3.0 in order to enable an easy transition from the "basic" style to LYX.

In total, this should get you started in no time.

A.2 STYLE OPTIONS

There are a couple of options for `classicthesis.sty` that allow for a bit of freedom concerning the layout:

. . . or your supervisor might use the margins for some comments of her own while reading

- General:
  - `drafting`: prints the date and time at the bottom of each page, so you always know which version you are dealing with. Might come in handy not to give your Prof. that old draft.
- Parts and Chapters:
  - `parts`: if you use Part divisions for your document, you should choose this option. (Cannot be used together with `nochapters`.)
  - `linedheaders`: changes the look of the chapter headings a bit by adding a horizontal line above the chapter title. The chapter number will also be moved to the top of the page, above the chapter title.
- Typography:
  - `eulerchapternumbers`: use figures from Hermann Zapf's Euler math font for the chapter numbers. By default, old style figures from the Palatino font are used.
  - `beramono`: loads Bera Mono as typewriter font. (Default setting is using the standard CM typewriter font.)
  - `eulermath`: loads the awesome Euler fonts for math. Palatino is used as default font.

Options are enabled via `option=true`

- Table of Contents:
  - `tocaligned`: aligns the whole table of contents on the left side. Some people like that, some don't.
  - `dottedtoc`: sets pagenumbers flushed right in the table of contents.
  - `manychapters`: if you need more than nine chapters for your document, you might not be happy with the spacing between the chapter number and the chapter title in the Table of Contents. This option allows for additional space in this context. However, it does not look as "perfect" if you use `\parts` for structuring your document.
- Floats:
  - `listings`: loads the `listings` package (if not already done) and configures the List of Listings accordingly.
  - `floatperchapter`: activates numbering per chapter for all floats such as figures, tables, and listings (if used).

Furthermore, pre-defined margins for different paper sizes are available, e. g., `a4paper`, `a5paper`, and `letterpaper`. These are based on your chosen option of `\documentclass`.

The best way to figure these options out is to try the different possibilities and see what you and your supervisor like best.

In order to make things easier, classicthesis-config.tex contains some useful commands that might help you.

A.3 CUSTOMIZATION

This section will show you some hints how to adapt classicthesis to your needs.

The file classicthesis.sty contains the core functionality of the style and in most cases will be left intact, whereas the file classicthesis-config.tex is used for some common user customizations.

The first customization you are about to make is to alter the document title, author name, and other thesis details. In order to do this, replace the data in the following lines of classicthesis-config.tex:

Modifications in classic-thesis-config.t

```
% ****************************************************
% 2. Personal data and user ad-hoc commands
% ****************************************************
\newcommand{\myTitle}{A Classic Thesis Style\xspace}
\newcommand{\mySubtitle}{An Homage to...\xspace}
```

Further customization can be made in classicthesis-config.tex by choosing the options to classicthesis.sty (see Section A.2) in a line that looks like this:

```
\PassOptionsToPackage{
  drafting=true,    % print version information on the bottom
     of the pages
  tocaligned=false, % the left column of the toc will be
    aligned (no indentation)
  dottedtoc=false,  % page numbers in ToC flushed right
  parts=true,       % use part division
  eulerchapternumbers=true, % use AMS Euler for chapter font
    (otherwise Palatino)
  linedheaders=false,       % chaper headers will have line
    above and beneath
  floatperchapter=true,     % numbering per chapter for all
    floats (i.e., Figure 1.1)
  listings=true,    % load listings package and setup LoL
  subfig=true,      % setup for preloaded subfig package
  eulermath=false,  % use awesome Euler fonts for
    mathematical formulae (only with pdfLaTeX)
  beramono=true,    % toggle a nice monospaced font (w/ bold)
  minionpro=false   % setup for minion pro font; use minion
    pro small caps as well (only with pdfLaTeX)
}{classicthesis}
```

Many other customizations in classicthesis-config.tex are possible, but you should be careful making changes there, since some changes could cause errors.

A.4 ISSUES

This section will list some information about problems using classicthesis in general or using it with other packages.

Beta versions of classicthesis can be found at Bitbucket:

https://bitbucket.org/amiede/classicthesis/

There, you can also post serious bugs and problems you encounter.

A.5 FUTURE WORK

So far, this is a quite stable version that served a couple of people well during their thesis time. However, some things are still not as they should be. Proper documentation in the standard format is still missing. In the long run, the style should probably be published separately, with the template bundle being only an application of the style. Alas, there is no time for that at the moment… it could be a nice task for a small group of LaTeXnicians.

Please do not send me email with questions concerning LaTeX or the template, as I do not have time for an answer. But if you have comments, suggestions, or improvements for the style or the template in general, do not hesitate to write them on that postcard of yours.

A.6 BEYOND A THESIS

The layout of `classicthesis.sty` can be easily used without the framework of this template. A few examples where it was used to typeset an article, a book or a curriculum vitae can be found in the folder `Examples`. The examples have been tested with `latex` and `pdflatex` and are easy to compile. To encourage you even more, PDFs built from the sources can be found in the same folder.

A.7 LICENSE

GNU GENERAL PUBLIC LICENSE: This program is free software; you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation; either version 2 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but *without any warranty*; without even the implied warranty of *merchantability* or *fitness for a particular purpose*. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program; see the file `COPYING`. If not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston, MA 02111-1307, USA.

CLASSICHTHESIS AUTHORS' NOTE: There have been some discussions about the GPL's implications on using `classicthesis` for theses etc. Details can be found here:

https://bitbucket.org/amiede/classicthesis/issues/123/

We chose (and currently stick with) the GPL because we would not like to compete with proprietary modified versions of our own work. However, the whole template is free as free beer and free speech. We will not demand the sources for theses, books, CVs, etc. that were created using `classicthesis`.

Postcards are still highly appreciated.

BIBLIOGRAPHY

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. "Concrete problems in AI safety." In: *arXiv preprint arXiv:1606.06565* (2016).

[2] Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. "Coherent measures of risk." In: *Mathematical finance* 9.3 (1999), pp. 203–228.

[3] Nicole Bäuerle and Jonathan Ott. "Markov decision processes with average-value-at-risk criteria." In: *Mathematical Methods of Operations Research* 74.3 (2011), pp. 361–379.

[4] Marc G Bellemare, Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." In: *arXiv preprint arXiv:1707.06887* (2017).

[5] Richard Bellman. "A Markovian decision process." In: *Journal of Mathematics and Mechanics* (1957), pp. 679–684.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[7] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. "Risk-sensitive and robust decision-making: a CVaR optimization approach." In: *Advances in Neural Information Processing Systems.* 2015, pp. 1522–1530.

[8] Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. "Distributional Reinforcement Learning with Quantile Regression." In: *arXiv preprint arXiv:1710.10044* (2017).

[9] Javier Garcıa and Fernando Fernández. "A comprehensive survey on safe reinforcement learning." In: *Journal of Machine Learning Research* 16.1 (2015), pp. 1437–1480.

[10] Roger Koenker and Kevin F Hallock. "Quantile regression." In: *Journal of economic perspectives* 15.4 (2001), pp. 143–156.

[11] Erwin Kreyszig. *Introductory functional analysis with applications.* Vol. 1. wiley New York, 1989.

[12] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. "AI Safety Gridworlds." In: *arXiv preprint arXiv:1711.09883* (2017).

[13] Anirudha Majumdar and Marco Pavone. "How Should a Robot Assess Risk? Towards an Axiomatic Theory of Risk in Robotics." In: *arXiv preprint arXiv:1710.11040* (2017).

[14] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. "Nonparametric return distribution approximation for reinforcement learning." In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10).* 2010, pp. 799–806.

[15] Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. "Parametric return density estimation for reinforcement learning." In: *arXiv preprint arXiv:1203.3497* (2012).

[16] Georg Ch Pflug and Alois Pichler. "Time-consistent decisions and temporal decomposition of coherent risk functionals." In: *Mathematics of Operations Research* 41.2 (2016), pp. 682–699.

[17] R Tyrrell Rockafellar and Stanislav Uryasev. "Optimization of conditional value-at-risk." In: *Journal of risk* 2 (2000), pp. 21–42.

[18] R Tyrrell Rockafellar and Stanislav Uryasev. "Conditional value-at-risk for general loss distributions." In: *Journal of banking & finance* 26.7 (2002), pp. 1443–1471.

[19] Alexander Shapiro. "On Kusuoka representation of law invariant risk measures." In: *Mathematics of Operations Research* 38.1 (2013), pp. 142–152.

[20] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge, 1998.

[21] Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. "Policy gradient for coherent risk measures." In: *Advances in Neural Information Processing Systems*. 2015, pp. 1468–1476.

[22] John N Tsitsiklis. "Asynchronous stochastic approximation and Q-learning." In: *Machine learning* 16.3 (1994), pp. 185–202.

[23] Christopher JCH Watkins and Peter Dayan. "Q-learning." In: *Machine learning* 8.3-4 (1992), pp. 279–292.

DECLARATION

Put your declaration here.

Prague, January, 2018

Silvestr Stanko

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Thank you very much for your feedback and contribution.