Background

General motivations and goals

As a composer, my motivation is the *automation of creative processes* (AoCP). Here, I mean automation in the sense of "letting a machine perform the task". I consider it to be very relevant not only for obvious economic/social reasons, which apply to every kind of automation: it can also help to bring some new knowledge about ourselves, since automating a task requires being able to describe it in a precise way, and creativity is a predominantly human skill. Furthermore, the connection between organic and mechanic processes has inspired many different aesthetical positions, which I also find interesting and attractive.

In this context, the goal would be to achieve a precise definition of creativity, or, conversely, to define algorithms whose output could be regarded as creative.

Defining creativity

Of this two tasks, the former seems very unfeasable, not only because creativity is seen as a very complex cognitive task, and requires a lot of knowledge on cognition that we still don't have: it would also require a general consensus on its purely idiomatic usage/meaning, which would be very difficult to achieve due to the many social factors involved.

From my personal experience and discussions, the consensus gets as far as to say that something is regarded as creative if and only if it represents or produces some qualitative gain of knowledge on the receptor, that is: motivates the perception and comprehension of a new concept or category. And even this has some caveats:

— I believe that *qualitative* and *quantitative* differences are related, for example: overdriving the quantities of an existing setup can lead to the perception of new qualities. In any case, this is bound to the way of how do humans categorize, which is a very complex topic, and a matter of investigation itself.

— It suffices if the communicated concepts are *new* just to the receptor. Of course, the more global the audience, the safer is to say that something is totally new, but it can't be safely assumed that objective originality exists without going into deep metaphysical discussion.

An empirical approach

In this terms, the task of achieving a precise definition of creativity seems unattractive to me. In comparation, defining algorithms whose output could be regarded as creative seems much more feasable: it is much more localized, which can help to reach such

a consensus, and technically plausible: first develop the algorithms, guided by some heuristic (e.g. intuition or educated guess), and then speculate on how the results could fit/disregard the premises. But this empirical approach presents two major drawbacks:

— If it is too localized, won't be regarded as creative at all: a delay guitar pedal, for example, produces new sounds every time it is used, but a receptor would very unlikely become some categorical new knowledge as a result of the work of the device, since they are a mere reproduction of its input.

— As every other empirical approach, it relies heavily on some kind of belief, if there aren't empirical results to show, or at best convenience, if the results are good.

So the bottom-up, empirical strategy is to try to get a broad and sound technical basis, in order to be able to conceive models that generalize well, and search for convenient/hopeful results. This sounds very harsh: in the case of AoCP, even the criterium for measuring the results as good or bad is very weak. Still, it may work well, and there are aspects that haven't been discussed yet: here is where the aesthetical dimension comes to play.

My idea of art (incomplete section)

I regard my activity as an artistic one, because art is, for me, *any form of discourse that represents or communicates an idea in a way that involves an aesthetical positioning.* In my case, the *idea* is the AoCP, and the *positioning* is, for each case, the one that achieves an optimal representation and/or communication of the idea. This doesn't mean the aesthetical dimension is irrelevant, rather the opposite: the aesthetical position of each project must be such that the corresponding idea is "optimally" represented and/or transmitted. Optimal, for me, means:

— Free as in freedom: share of data, reproducibility but not necessarily decentralisation. machines are very advanced and there seems to be little awareness about that. Whatever we go for, better done with awareness.

— Complex but Global: popular contents to spread elitist means. Make elite grow "igualar por arriba": does this mean educational/pedagogic?

— Generalizability: I make music just bc i know more about that than other things. Unless we have a special soul or something, everything is a stream of patterns

In this sense, it has many common points with the hacker, pop, scientific discourses, and I anticipate many of their aesthetical components, but also many differences: popularity not only through simplicity! "autotune" involves heavy math but its very popular... TODO. There are also many common points with science... elaborate. Also religion, i talk about faith
For example, a common *topos* for artworks that tackle the AoCP is the convergence between human and machine: since there is no record of such a thing, many details

are open: the artist does have to specify how could it happen: do machines become a copy of the 19th century western adult, or do humans also get closer to machines by, for instance, abandoning the idea of subject and freedom?

In this context, my discourse doesn't only intend to be a pure empirical investigation on the sources of creativity, or a way to open up new ways for the industry: art can be a very effective way to transmite the belief that the AoCP is possible, and to create a starting point for discussion about our present and future among machines: and this not only as a mean, but also a goal itself.

The Wave-To-Wave Algorithm

Approach

Following the program explained in the background section, the basic thesis behind this project is that **the task of optimizing could be regarded as creative**. In Mathematics, optimization problems try to find the maximum (or minimum) of a function. If the function to optimize is convex, this means it has a single global maximum, which can be efficiently found most of the times following an analytical approach.
Non-convex functions, on the other side, present many local maxima, and the optimization problem is also non-convex if there is no general way to know where the global maximum is. The general solution would be here to search through the whole input space, which usually takes too long to be plausible. In that case it is necessary to define some **heuristics**, that is, criteria that help to reduce the search space but still reach a satisfying solution.

There are many kinds of optimization problems, and many of them can take place in a creative context. The problem corresponding to this algorithm is the following: given an audio file on one side, and a finite set of audio files on the other (the materials), **approximate the original file by transforming and superimposing the materials**. This approximation has to be so good that the original is clearly noticeable (in ideal cases, even speech should be understood), but the building blocks are perceived too.

There are some precedents in performing such a task, most notably IRCAM's *CataRT*, which usually rely on already granular materials, and process the result sequentially through time. The approach of the algorithm presented here is substantially different, since it is divided in two phases:

1. Provide the closest result between the original signal and the reconstruction, by minimizing their "energy difference". The energy difference between to signals is a convex function and its global minimum can be efficiently calculated, as it wil be shown. Keep in mind that finding the closest reconstruction possible, doesn't mean that it has to be good: unlike other models, this makes no assumptions on the inputs (apart from them bein discrete, real-valued signals).

2. Alter the setup to force suboptimalities in diverse ways.

In other words, instead of defining the approximating algorithm with implicit heuristics already from scratch, an optimum is calculated, and the heuristics are defined in comparison to it. The hope is that this helps to make them more explicit and therefore improve not only the quality of the compositional process, but also its results and their interpretation. With this approach, the first stage implies more or less the development of a "compositional tool": it is in the second stage where the above mentioned non-convexity is expected to happen.

Optimization objective

Given:

$$S_{(t)} \hat{=} \text{ a real-valued signal of sample length } |S| \in \mathbb{N} \text{ (the } \textit{original}\text{)}$$
$$\vec{m} := (\ m_{1(t)}, ..., m_{N(t)}) \hat{=} \text{ a vector of } N \in \mathbb{N} \text{ discrete, real-valued signals (the } \textit{materials}\text{)}$$
$$f(x_{(t)}) = k\varphi_{x(t+d)} \hat{=} \text{ a signal consisting of any non-linear transformation } \varphi_x \text{ and a further delay } d \in \mathbb{N} \text{ and normalization } k \in \mathbb{R}$$

Find the parameters for the signal $S'_{(t)}$ (the *reconstruction*):

$$S'_{(t)} = \sum_{n=1}^{N} \{f_n(m_{n(t)})\} = \sum_{n=1}^{N} \{k_n \varphi_{m_n(t+d_n)}\}$$

That minimize the energy of the *residual* signal $R_{(t)} := E[S_{(t)} - S'_{(t)}]$, that is, the difference between *original* and *reconstruction*. The energy of a signal is defined as:

$$E[x_{(t)}] = \sum_{t=1}^{|x|} \{(x_{(t)})^2\}$$

And therefore, the optimization objective remains then as:

$$(\vec{k}^*, \vec{d}^*) = \underset{\vec{k},\vec{d}}{\text{minimize}} \left[R_{(t)} = E[S_{(t)} - S'_{(t)}] = \sum_{t=1}^{|S|} \left\{ \left(S_{(t)} - \sum_{n=1}^{N} \{k_n \varphi_{m_n(t+d_n)}\} \right] \right.$$

Which can also be written in the following equivalent algebraic expression:

$$(\vec{k}^*, \vec{d}^*) = \underset{\vec{k},\vec{d}}{\text{minimize}} \left\| \vec{S} - \Phi_{\vec{d}} \vec{k} \right\|_2^2$$

With the Signals and parameters defined as:

$$\vec{S} = \begin{pmatrix} S_{(1)} \\ \vdots \\ S_{(|S|)} \end{pmatrix} \quad \vec{k} = \begin{pmatrix} k_1 \\ \vdots \\ k_N \end{pmatrix} \quad \vec{d} = \begin{pmatrix} d_1 \\ \vdots \\ d_N \end{pmatrix} \quad \Phi_{\vec{d}} = \begin{pmatrix} \varphi_1(m_{1(1+d_1)}) & \cdots & \varphi_N(m_{N(1+d_N)}) \\ \varphi_1(m_{1(2+d_1)}) & \cdots & \varphi_N(m_{N(2+d_N)}) \\ \vdots & \ddots & \vdots \\ \varphi_1(m_{1(|S'|)}) & \cdots & \varphi_N(m_{N(|S'|)}) \end{pmatrix}$$

Whereas the φ signals inside Φ are zero-padded (filled up with zeroes) in order to be defined between 1 and $|S'|$. As it can be seen in the proof, the solutions for the minimization problem have the following values:

$$\vec{d}_n^* = ??$$
$$\vec{k}_n^* = ??$$

Whereas $CC[S_{(t)}, \varphi_{n(t)}]$ is the *cross-correlation* between signals $S_{(t)}$ and $\varphi_{n(t)}$.

Proof for the optimization objective and algorithm

The minimization of $E[R_{(t)}] = \left\|\vec{S} - \Phi_{\vec{d}}\vec{k}\right\|_2^2$ for k corresponds to the well-known problem of the *least squares*, solvable with the *normal equations*. For any given $\vec{k}$ and $\vec{d}$, the derivative of the corresponding direction δk is:

$$\nabla_{\vec{k}}\left\|\vec{S} - \Phi_{\vec{d}}\vec{k}\right\|_2^2 \cdot \delta\vec{k} = 2\langle \Phi_{\vec{d}}\delta\vec{k}\ ,\ \vec{S} - \Phi_{\vec{d}}\vec{k}\rangle = 2\delta\vec{k}^T(\Phi_{\vec{d}}^T\vec{S} - \Phi_{\vec{d}}^T\Phi_{\vec{d}}\vec{k})$$

In this context, for a fixed $\vec{d}$ the minimum of $\vec{k}$ occurs when this derivative is set to zero, which has a single solution at $\Phi_{\vec{d}}^T\vec{S} - \Phi_{\vec{d}}^T\Phi_{\vec{d}}\vec{k}$:

$$\vec{k}_{opt} = (\Phi_{\vec{d}}^T\Phi_{\vec{d}})^{-1}\Phi_{\vec{d}}^T\vec{S}$$

But as it can be observed, the value of $\vec{k}_{opt}$ is also affected by changes in $\vec{d}$. Hence, the optimal value $\vec{d}^*$ is defined as the one that provides the best $\vec{k}_{opt}$ possible, that is $\vec{k}^*$:

$$\vec{k}^* = (\Phi_{\vec{d}^*}^T\ \Phi_{\vec{d}^*})^{-1}\Phi_{\vec{d}^*}^T\ \vec{S} = \Phi_{\vec{d}^*}^+\ \vec{S}$$

Whereas $\Phi_{\vec{d}}^+ = (\Phi_{\vec{d}}^T\Phi_{\vec{d}})^{-1}\Phi_{\vec{d}}^T$ is the **Moore-Penrose pseudoinverse** of $\Phi_{\vec{d}}$.
Recovering the original minimization problem in its algebraic form,

$$(\vec{k}^*, \vec{d}^*) = \underset{\vec{k},\vec{d}}{\text{minimize}}\ \left\|\vec{S} - \Phi_{\vec{d}}\vec{k}\right\|_2^2$$

It is now possible to substitute $\vec{k}$ with the already known expression for $\vec{k}_{opt}$, and since it only depends on $\vec{d}$ and $\vec{S}$, the optimization problem gets simplified:

$$\vec{d}^* = \underset{\vec{d}}{\text{minimize}}\ \left\|\vec{S} - \Phi_{\vec{d}}(\Phi_{\vec{d}}^+\vec{S})\right\|_2^2 = \underset{\vec{d}}{\text{minimize}}\ \left\|\vec{S} - \Psi_{\vec{d}}\vec{S}\right\|_2^2$$

Whereas $\Psi_{\vec{d}} = \Phi_{\vec{d}}\Phi_{\vec{d}}^+$ is the **orthogonal projector** of $\Phi_{\vec{d}}$. This matrix has various valuable properties. Given $\Phi_{\vec{d}} \in \mathbb{R}^{a\times b}$, the following holds for $\Psi_{\vec{d}}$:

1. it is a square matrix ($\Psi_{\vec{d}} \in \mathbb{R}^{a\times a}$)

2. it is symmetric ($\Psi_{\vec{d}} = \Psi_{\vec{d}}^T$)
3. it is idempotent ($\Psi_{\vec{d}} = \Psi_{\vec{d}}^2$)
4. it has a eigenvalues, and from them $min(a, b)$ are ones and the rest zeros. The corresponding eigenspaces are, respectively, its range and kernel.
5. Self-Adjoint ($\Psi_{\vec{d}}\Phi_{\vec{d}} = \Phi_{\vec{d}}$, or, conversely, $\Psi_{\vec{d}} m_{i(t+d_i)} = m_{i(t+d_i)}$)
6. Provides a direct relation between any given original and its optimal reconstruction: ($\vec{S'^*} = \Psi_{\vec{d}^*}\vec{S}$).
7. For every two signals x, y, it holds: $\langle \Psi x, y\rangle = \langle x, \Psi y\rangle = \langle \Psi x, \Psi y\rangle$

Proofs for this are not included here, but can be found in [WIKIPEDIA STFG]. Using some of this properties, it is possible to further reformulate the optimization objective into a more convenient expression:

$$\begin{aligned}
\vec{d}^* &= \underset{\vec{d}}{\text{minimize}} \left\| \vec{S} - \Psi_{\vec{d}}\vec{S} \right\|_2^2 \\
&= \underset{\vec{d}}{\text{minimize}} \langle \vec{S} - \Psi_{\vec{d}}\vec{S}\ ,\ \vec{S} - \Psi_{\vec{d}}\vec{S}\rangle \\
&= \underset{\vec{d}}{\text{minimize}} \langle \vec{S}, \vec{S}\rangle + \langle \Psi_{\vec{d}}\vec{S}, \Psi_{\vec{d}}\vec{S}\rangle - 2\langle \vec{S}, \Psi_{\vec{d}}\vec{S}\rangle \\
&= \underset{\vec{d}}{\text{minimize}} \langle \Psi_{\vec{d}}\vec{S}, \Psi_{\vec{d}}\vec{S}\rangle - 2\langle \vec{S}, \Psi_{\vec{d}}\vec{S}\rangle \\
&= \underset{\vec{d}}{\text{minimize}} \vec{S}^T \Psi_{\vec{d}}^T \Psi_{\vec{d}}\vec{S} - 2\vec{S}^T \Psi_{\vec{d}}\vec{S} \\
&= \underset{\vec{d}}{\text{minimize}} \vec{S}^T \Psi_{\vec{d}}\vec{S} - 2\vec{S}^T \Psi_{\vec{d}}\vec{S} \\
&= \underset{\vec{d}}{\text{maximize}} \vec{S}^T \Psi_{\vec{d}}\vec{S} \\
&= \underset{\vec{d}}{\text{maximize}} \langle \vec{S}\ , \Psi_{\vec{d}}\vec{S}\rangle
\end{aligned}$$

And since $\vec{S}$ is given and not altered through the optimization process, it follows a spectral analysis of $\Psi_{\vec{d}}$ in order to understand better how the properties of $\Phi_{\vec{d}}$ and changes on $\vec{d}$ affect the outcome. First, a quick reminder on SVD decomposition: any matrix $M \in \mathbb{R}^{a\times b}$ can be decomposed in the following expression:

$$M = U\Sigma V^T$$

whereas:

$U \in \mathbb{R}^{a\times a}$ is an orthogonal matrix ($UU^T = U^TU = I$)

$V \in \mathbb{R}^{b\times b}$ is an orthogonal matrix ($VU^T = V^TV = I$)

$\Sigma \in \mathbb{R}^{a\times b}$ is a diagonal matrix containing the *singular values* of M

$M^T = V\Sigma U^T$ is the transpose

$M^+ = V\Sigma^+U^T$is the pseudoinverse (equivalent to the Moore-Penrose pseudoinverse)

Σ^+ is the pseudoinverse of Σ, and can be calculated by replacing each non-zero diagonal entry λ by its reciprocal $\frac{1}{\lambda}$, and transposing after.

With this knowledge, it is possible to express $\Psi_{\vec{d}}$ in terms of the SVD decomposition of $\Phi_{\vec{d}} = U\Sigma V^T$:

$$\begin{aligned}\Psi_{\vec{d}} &= \Phi_{\vec{d}}\Phi^+_{\vec{d}} \\ &= U\Sigma V^T V\Sigma^+U^T \\ &= U\Sigma\Sigma^+U^T\end{aligned}$$

Which, combined with the properties of U and Σ, leads to the following conclusions:

— For a given $\Phi_{\vec{d}} \in \mathbb{R}^{a\times b}$ with b linearly independent columns, the diagonal matrix $\Sigma\Sigma^+ \in \mathbb{R}^{b\times b}$ will have a total of a diagonal entries. From this total, $min(a, b)$ will equal to one, and the rest will equal to zero. This leads to the following observation:

$$\begin{aligned} b \geq a &\iff \Sigma\Sigma^+ = I \in \mathbb{R}^{a\times a} \\ &\iff \Psi_{\vec{d}} = U\Sigma\Sigma^+U^T = UIU^T = UU^T = I \in \mathbb{R}^{a\times a} \\ &\iff \vec{S}' = \Psi_{\vec{d}}\vec{S} = I\vec{S} = \vec{S} \\ &\iff \vec{S}' = \vec{S}\end{aligned}$$

In other words:

> If the original signal has a length of a samples, and the set of materials has at least a linearly independent elements, it is possible to fully reconstruct the original as a linear combination of a-many of such elements

— The diagonal matrix $\Sigma\Sigma^+$ acts like a filter: when it is totally open, it forms an identity matrix that leads to no transformation of the input to $\Psi_{\vec{d}}$ (because $UU^T = I$). When an entry is zero, the corresponding dimension will be filtered out.

> The best case of maximizing $\langle \vec{S}', \Psi_{\vec{d}} \vec{S} \rangle$ is achieved when $\vec{S}' = \vec{S}$, because orthogonal matrices preserves the length

will be the maximum value of the optimization objective (namely to maximize $\langle \vec{S}', \Psi_{\vec{d}} \vec{S} \rangle$), is achieved when $\vec{S}' = \vec{S}$, because orthogonal matrices preserve the length UU^T

This case would be uninteresting for the described setup (the)

.

Now, performing a SVD decomposition of the psi matrix shows that it has exactly p=min(col, row) non-zero singular values, and that all of them are equal to 1. This means many things, one ofthem is that PSI ALWAYS SHRINKS THE LENGTH OF ITS INPUT. In words, finding the $\vec{d}$ and $\vec{k}$ that minimize the energy between $\vec{S}$ and $\vec{S}'$ is equivalent to find the $\vec{d}$

STRATEGY TO CONTINUE: SHOW THAT THIS IS EQUIVALENT TO MINIMIZING THE FROBENIUS BETWEEN PSI AND I... MAYBE CEILING ANALYZE OF WORST-CASES ALLOWS AN EXPLICIT FORMULATION OF THE DIFFERENT BETWEEN PSI AND I, AND MINIMIZING THIS FORMULA MIGHT BE A CONVEX PROBLEM (IN SOME HIGHER DIMENSIONAL SPACE MAYBE).

Analysis of the model

The whole problem can be divided into three smaller problems, clearly differentiated by their complexity and how they interact with each other. The thinking process leading to this division goes as follows:

1. This model avoids intentionally any kind of limitations on S and M. Therefore, no assumptions can be done on them, further than the fact that they are real-valued, discrete signals. This is done so in order to keep the applications of this algorithm as broad as possible, but this also implies that the input space is huge, and leaves the selection of the input signal as an open problem. This is intended to be so, the algorithm will assume here that some S and M are given.

2. The model also avoids any kind of limitations on the $\vec{\varphi} = \{\varphi_1, ..., \varphi_N\}$ transformations, even if they are highly non-linear or even non-deterministic. Also, this enhances a lot its generality, but increases even more its own complexity, and the complexity of its output. In order to keep the simplicity of the model, it will also be assumed that the space of possible transformations is given to the algorithm, and not calculated. To do that, it suffices to define all transformed materials as given materials, and incorporate them to the M set. We define this way the M' set:

$$M' := M \cup \{\varphi_i[m_i] \quad \forall\, \varphi_i[m_i] \in M\}$$

This has two main drawbacks: the size of the input space becomes even bigger, and the transformations have to belong to the preprocessing stage. But the advantage is that model remains linear:

$$S'_{(t)} = \sum_{n=1}^{N} \{k_n \cdot m_{n(t+d_n)}\} \qquad : \quad \forall\, m_{n(t)} \in M'$$

Which is very convenient, since it can be highly optimized as it will be explained. The hope is that this optimization makes up for the problem of an oversized input space and preprocessing stage. Of course, further optimizations can and will be done, but they imply the reduction of the input and transformation spaces in some way, and are therefore kept outside of the model: the intention here is to formulate it as general as possible.

Numbers

Since the problem is a linear one, the opt. objective has some nice properties:
max(CC(S, lincomb(m))) = lincomb(max(CC(s,m))) because associativity
min E... big numbers!!
This proximity has been chosen for two reasons: the ideal deal And the *proximity* between S and S' is defined as:

$$J(S, S') = ???$$

The objective is to maximize J. The delays are trivial (show CC, distribut. property).
The k$_s$*arealittlemoretricky*(*showthatdisindependentfromk*).*Showminimizationofenergy*
Starting from the sum-of-squares error function

$$E_D(w) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\}^2$$

derive the maximum likelihood solution for the parameters

$$w_{ML} = (\Phi^T \Phi)^{-1} \Phi^T t$$

where

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & ... & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & ... & \phi_{M-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & ... & \phi_{M-1}(x_N) \end{pmatrix}$$

is the design matrix with basis functions $\phi_j(x_i)$, $X = \{x_1, ...x_n\}$ the vectors of input training data and $t = \{t_1, ...t_n\}$ corresponding output training values.

i)

A very clear and brief explanation to this topic can be found in the *Deep Learning Book*, page 109: Since the energy function is quadratic, the optimization solution for one dimension is convex and can be anallitically found by equalling the derivative to zero. But furthermore, a linear combination of convex optimization problems is itself convex too, since multiplication by a scalar and addition of linearly independent terms doesn't affect the outcome: the multiplicating scalars can be ignored (In fact, the usual normalization factor $\frac{1}{M}$ was here disregarded), and the different problems optimized separately.
This is the case when performing parametric multivariate linear regression, with the L_2 energy function: each weight is linearly independent from all others, and contributes to E in a convex way.

ii)

In this terms the optimization objective can be fo rmuled as follows:

$$w_{ML} = \underset{w}{\text{minimize}} \quad E(w, \Phi, t) = \|\Phi w - t\|_2^2 = (\Phi w - t)^T(\Phi w - t)$$

And the analytical way to find it:

$$\begin{aligned}
\frac{\partial}{\partial w_{ML}} E(w_{ML}, \Phi, t) = 0 &\iff \frac{\partial}{\partial w_{ML}} \Big((\Phi w_{ML} - t)^T (\Phi w_{ML} - t) \Big) = 0 \\
&\iff \frac{\partial}{\partial w_{ML}} \Big(w_{ML}^T \Phi^T \Phi w_{ML} - 2 w_{ML}^T \Phi^T t + t^T t \Big) = 0 \\
&\implies 2\Phi^T \Phi w_{ML} - 2 w_{ML}^T \Phi^T t = 0 \\
&\implies w_{ML} = (\Phi^T \Phi)^{-1} \Phi^T t
\end{aligned}$$

□

Exercise 2

Consider a data set in which each data point (x_n, t_n) is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(w) = \frac{1}{2} \sum_n r_n (t_n - w^T \phi(x_n))^2$$

Find an expression for the solution $w*$ that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise variance and (ii) replicated data points.

i)

The energy function can be reformulated as follows:

$$\frac{1}{2}\sum_n r_n(t_n - w^T\phi(x_n))^2 = \frac{1}{2}\sum_n {}_+\sqrt{r_n}^2(t_n - w^T\phi(x_n))^2$$
$$= \frac{1}{2}\sum_n ({}_+\sqrt{r_n}t_n - {}_+\sqrt{r_n}w^T\phi(x_n))^2$$

Which brings up the very same optimization objective as the one shown in Exercise 1 (assuming that the weighting factors are given and not learned, that is). Therefore, just two pre-processing calculations are needed:

$$t_{ML} = t \circ {}_+\sqrt{r}$$
$$\Phi_{ML} = \Phi \odot {}_+\sqrt{r}$$

Whereas ${}_+\sqrt{r}$ is the element-weise positive square root of the r vector in $\mathbb{R}^N$, $a \circ b$ represents the element-weise multiplication of two vectors of same dimensionality, and $a \odot b$ abuses this notation, to represent the element-weise multiplication of each vector in the matrix a with the vector b. In This terms, the solutions is simply to apply the normal equations to the preprocessed data:

$$w_{ML} = (\Phi_{ML}^T\Phi_{ML})^{-1}\Phi_{ML}^T t_{ML}$$

□

Exercise 3

Generate own data sets, e.g. using $t = f(x)+0.2\epsilon$ *with* $f(x) = sin(2\pi x)$ *and* $\epsilon \sim \mathcal{N}(0,1)$*, and illustrate the bias-variance decomposition by fitting a polynomial model* $y(x;w) = \sum_{i=0}^{r} w^i x^r$ *to many different data sets* $D_1, ..., D_L$ *, each of length* N*.*
Let w^D *denote the parameters minimizing the mean squared error on dataset* D*. Then,*

$$bias^2 \approx \frac{1}{L}\sum_l \frac{1}{N}\sum_n (\overline{y}(x) - f(x))^2$$
$$variance \approx \frac{1}{L}\sum_l \frac{1}{N}\sum_n (y(x; w*^{D_t}) - \overline{y}(x)^2$$

where $\overline{y}(x) = \frac{1}{L}\sum_t y(x; w*^{D_t})$

Solution:

See/execute Python2 script*fernandez_blatt5.py* for the details. As explained in the *lecture's slides*), the L_2 loss function can be decomposed into **bias**^{2} **+ variance + noise**. For a given dataset of limited size, only limited assumptions can be done: if only the variance term is taken into account (that is, no regularization term is provided), the hypothesis will maximize its adaptation to the dataset, potentially fitting perfectly to it, but it will fail to generalize, that is: to capture the underlying features. On the other side, a model excesively based on the bias term (that is, with a very high regularization index), will penalize every hypothesis that goes too far away from some given assumptions (in this case, the overall distance to the zero-vector). This assumptions may be unrealistic and relying heavily on them may be therefore a bad strategy.
Both terms are based on the same input parameters, but represent opposite ideas. The bottom line behind this explanation is that a **bias-variance tradeoff** takes always place for datasets of limited size, and, unless the size of the dataset can be increased, a compromise between both of them must be achieved. This is typically achieved by testing many different regularization factors, and cross-validating the results, as shown in the Figure 1:

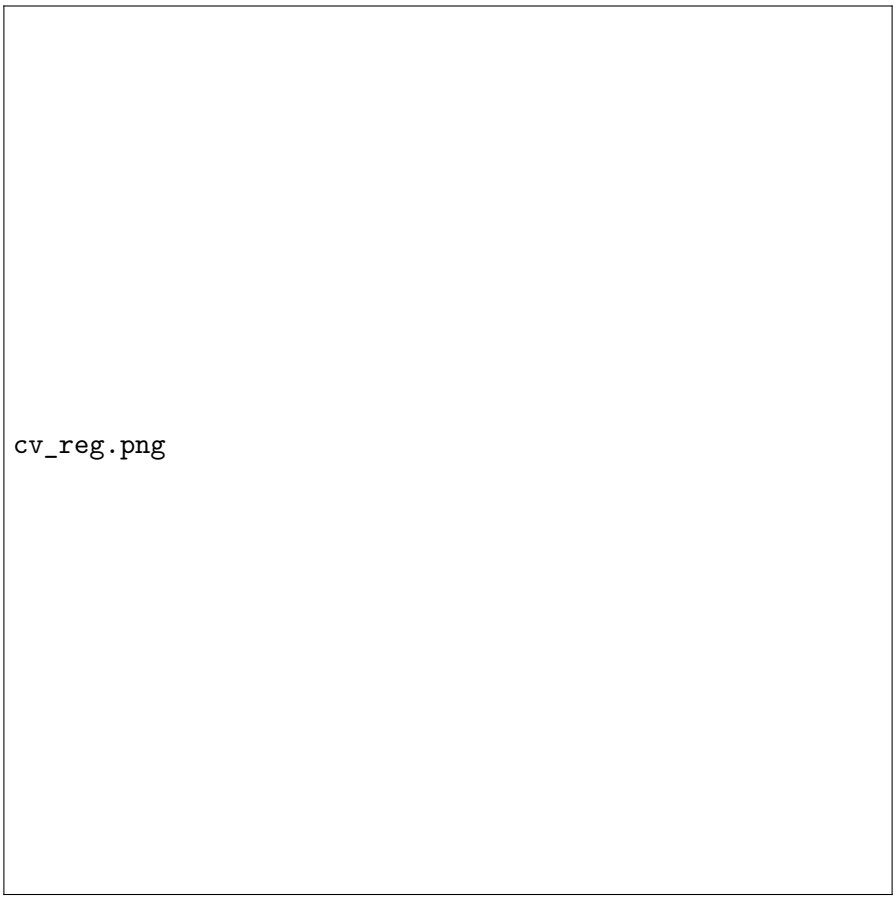

Figure 1: test error and its decomposition for different reg. factors (lecture slides)

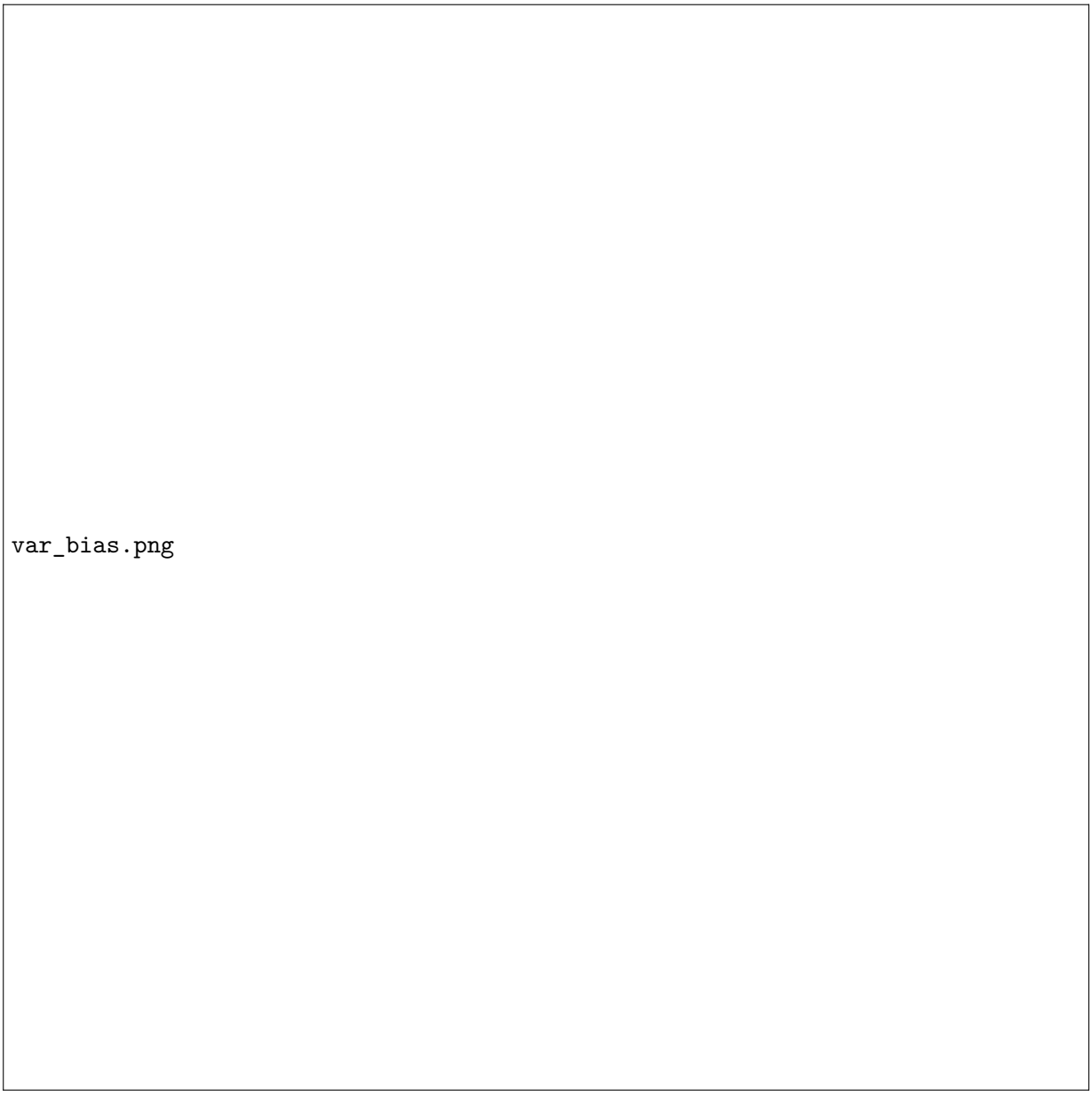

Figure 2: generated example illustrating a case of variance (blue) vs. bias (green) tradeoff