Concrete MAP Detection: A Machine Learning Inspired Relaxation

Edgar Beck, Carsten Bockelmann and Armin Dekorsy
Department of Communications Engineering
University of Bremen, Bremen, Germany
Email: {beck, bockelmann, dekorsy}@ant.uni-bremen.de

***Abstract*—Motivated by large linear inverse problems where the complexity of the Maximum A-Posteriori (MAP) detector grows exponentially with system dimensions, e.g., large MIMO, we introduce a method to relax a discrete MAP problem into a continuous one. The relaxation is inspired by recent ML research and offers many favorable properties reflecting its quality. Hereby, we derive an iterative detection algorithm based on gradient descent optimization: Concrete MAP Detection (CMD). We show numerical results of application in large MIMO systems that demonstrate superior performance w.r.t. all considered State of the Art approaches.**

***Index Terms*—MAP, maximum a-posteriori, Large MIMO detection, concrete distribution, Gumbel-Softmax, Machine Learning**

I. Introduction

After the Deep Learning revolution of the 2010s, Machine Learning (ML) recently gained a lot of attention from the digital signal processing community and the ”AI winter” has ended [1], [2]. It was just a matter of time until researchers wondered how communication systems can benefit from the new insights. The main advantage of ML lies in handling model and algorithm deficits.

To a great extent, a model deficit does not apply to wireless communications since it features well-established models, e.g., AWGN. These models describe reality well and enable development of optimized algorithms. However, these algorithms may be too complex to be implemented and an algorithm deficit results [1].

In large linear inverse problems, e.g., typical for large MIMO (Multiple Input Multiple Output) systems with many antennas at transmitter and receiver side [3], we have such an algorithm deficit. On the one hand, Maximum A-Posteriori (MAP) detection at the receiver exhibits high computational complexity growing exponentially with system dimensions. Even the efficient implementation, the sphere detector, remains too complex in such a scenario [4]. On the other hand, suboptimal solutions like linear detectors show bad performance.

Therefore, we propose a new detection/classification approach inspired by recent ML research: Concrete MAP Detection (CMD). We relax the discrete Random Variables (RVs) of the MAP problem by means of the continuous concrete distribution [5], [6]. It was recently discovered in the context of large scale stochastic computation graphs and enables differentiation through discrete stochastic nodes.

The proposed relaxation offers many favorable properties: On the one hand, the probability distribution function (pdf) of the relaxed continuous RVs converges to the exact pdf of the discrete RVs in the hyperparameter limit. Hence, also the expected values converge. On the other hand, we notice good algorithmic properties. First, a reparametrization of the relaxed RVs makes knowledge of the true pdf for solving the MAP problem irrelevant. Second, CMD allows to differentiate continuously through the MAP cost function in any non-linear probabilistic model. Finally, the hyperparameter can be, e.g., adapted, to improve algorithmics.

Furthermore, we show first numerical results of application in large MIMO systems demonstrating only small performance loss compared to discrete optimization.

II. Theoretical Background

A. System Model and Problem Statement

To motivate the concrete relaxation, we consider a linear observation model typically encountered in MIMO systems. Here, we assume $\mathbf{x}$ to be a normalized multivariate discrete Random Variable (RV), i.e., $\mathbf{x} = \{x_n\}_{n=1}^{N_\mathrm{T}}$ with $\mathbb{E}[|x_n|^2] = 1$, whose elements are from a real-valued set $\mathcal{M}$. The RV passes a linear channel $\mathbf{H} \in \mathbb{R}^{N_\mathrm{R} \times N_\mathrm{T}}$ with i.i.d. Gaussian taps $h_{mn} \sim \mathcal{N}(0, 1/N_\mathrm{R})$. Then, the resulting RV is superimposed by Gaussian noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \sigma_\mathrm{n}^2 \mathbf{I}_{N_\mathrm{R}})$ with variance σ_n^2 and observed:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}\,. \tag{1}$$

The matrix $\mathbf{I}_{N_\mathrm{R}}$ denotes the identity matrix of dimension $N_\mathrm{R} \times N_\mathrm{R}$. To decide on the discrete multivariate RV $\mathbf{x}$ given the linear observation $\mathbf{y} \in \mathbb{R}^{N_\mathrm{R} \times 1}$, we solve the MAP problem

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x} \in \mathcal{M}^{N_\mathrm{T} \times 1}} p(\mathbf{x}|\mathbf{y}) \tag{2a}$$

$$= \arg \max_{\mathbf{x} \in \mathcal{M}^{N_\mathrm{T} \times 1}} p(\mathbf{y}|\mathbf{x}) \cdot p(\mathbf{x}) \tag{2b}$$

$$= \arg \min_{\mathbf{x} \in \mathcal{M}^{N_\mathrm{T} \times 1}} -\ln p(\mathbf{y}|\mathbf{x}) - \ln p(\mathbf{x}) \tag{2c}$$

$$\text{with} \quad p(\mathbf{y}|\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^{N_\mathrm{R}} \sigma_\mathrm{n}^{2N_\mathrm{R}}}} e^{-\frac{1}{2\sigma_\mathrm{n}^2}(\mathbf{y}-\mathbf{H}\mathbf{x})^T(\mathbf{y}-\mathbf{H}\mathbf{x})} \tag{3}$$

as the conditional pdf and $p(\mathbf{x})$ as the a-priori pdf. Since $x_n \in \mathcal{M}$, we have to do an exhaustive search over all element combinations to solve the MAP problem. We mention the sphere detector as an efficient implementation [4]. For

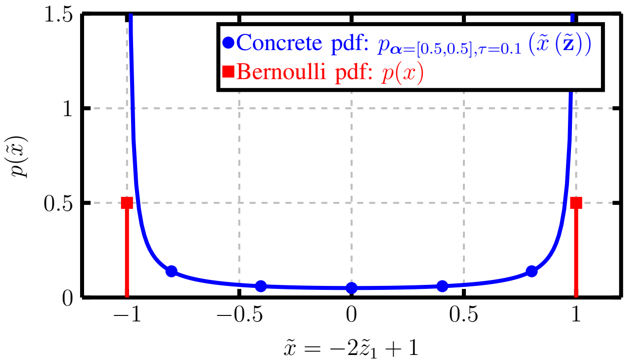

Fig. 1. The conrete pdf relaxes the Bernoulli distribution into the interior.

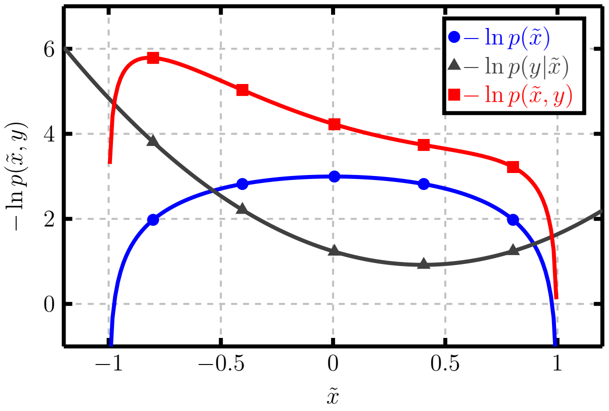

Fig. 2. Examplary plot of the conrete binary MAP cost function and the contribution of conditional and prior pdf to it.

complexity reduction in large systems, the discrete RV $\mathbf{x}$ is usually relaxed to be continuous. For example, we obtain the linear MMSE solution if we allow $\mathbf{x} \in \mathbb{R}^{N_\mathrm{T} \times 1}$ in (2c) to be continuous rather than discrete with $p(\mathbf{x}) = \mathcal{N}(\mathbf{0}_{N_\mathrm{T}}, \mathbf{I}_{N_\mathrm{T}})$ since mean and mode of Gaussian RVs coincide.

B. Concrete Distribution

In order to reduce the complexity of the MAP problem (2c), we propose the following CONtinuous relaxation of disCRETE variables: We replace the prior distribution $p(\mathbf{x})$ by means of the CONCRETE distribution or Gumbel-Softmax distribution recently discovered in the ML community [5], [6].

To explain the idea, let us assume that we have the discrete binary RV $x \in \mathcal{M}$ with $\mathcal{M} = \{-1, +1\}$. Further, let us define the discrete RV $\mathbf{z}$ as a one-hot vector where all elements are zero except for one element, i.e., $\mathbf{z} \in \{0, 1\}^{2\times 1}$ with values $\mathbf{z}_1 = [1, 0]^T$, $\mathbf{z}_2 = [0, 1]^T$. In addition, we collect the values of $\mathcal{M}$ in the representer vector $\mathbf{m} = [-1, 1]^T$. That way, we can write $x = \mathbf{z}^T\mathbf{m}$, e.g., $x = [1, 0] \cdot [-1, 1]^T = -1$.

We point out that the one-hot vector $\mathbf{z} \in \{0, 1\}^{M\times 1}$ represents a categorical RV with $M = |\mathcal{M}|$ classes. Motivated by classification [5], [6], the one-hot vector can be defined to be:

$$\mathbf{z} = \text{one-hot}\left(\arg\max_j \left[\mathbf{g} + \ln \boldsymbol{\alpha}\right]\right) . \tag{4}$$

Note that $\mathbf{g} \in \mathbb{R}^{M\times 1}$ is a multivariate continuous RV whose elements are i.i.d. Gumbel distributed while $\boldsymbol{\alpha} \in (0, 1)^{M\times 1}$ is the vector of class probabilities that sum up to one.

So far, x is discrete. The idea is to replace the $\arg\max$ computation of the so-called Gumbel-Max trick (4) by the softmax function

$$\tilde{\mathbf{z}} = f\left(\mathbf{g}\right) = \frac{e^{(\ln(\boldsymbol{\alpha})+\mathbf{g})/\tau}}{\sum_{i=1}^{M} e^{(\ln \alpha_i + g_i)/\tau}} . \tag{5}$$

The RV $\tilde{\mathbf{z}} \in (0, 1)^{M\times 1}$ is the so called concrete or Gumbel-Softmax RV and now continuous, e.g., $\tilde{\mathbf{z}} = [0.2, 0.8]$. It is controlled by a hyperparameter, the softmax temperature τ. The distribution of $\tilde{\mathbf{z}}$ in (5) was found to have a closed form density which is taken to be the definiton of the concrete distribution:

$$p_{\boldsymbol{\alpha},\tau}(\tilde{\mathbf{z}}) = (M-1)!\; \tau^{M-1} \prod_{k=1}^{M} \left(\frac{\alpha_k \tilde{z}_k^{-\tau-1}}{\sum_{i=1}^{M} \alpha_i \tilde{z}_i^{-\tau}} \right) . \tag{6}$$

With $\tilde{\mathbf{z}}$, we can also relax the discrete RV x into a continuous RV $\tilde{x}$ by defining $\tilde{x} = \tilde{\mathbf{z}}^T\mathbf{m}$. In Fig. 1, we illustrate the distribution $p(\tilde{x})$ for the special case $M = 2$ in comparison to the original categorical pdf $p(x)$ reducing to a Bernoulli pdf. It has the following properties reflecting the correctness of the relaxation [5]: First, we can reparametrize the concrete RV $\tilde{\mathbf{z}}$ and hence the RVs $\tilde{x}$ by Gumbel variables $\mathbf{g}$, a direct result from the initial idea (5). Moreover, rounding $\tilde{\mathbf{z}}$ restores a categorical variable. The same is true for the zero temperature limit $\tau \to 0$: The smaller τ, the more $\tilde{\mathbf{z}}$ approaches a categorical distribution and the approximation becomes more accurate. Thus, the statistics of x and $\tilde{x}$ remain the same for $\tau \to 0$.

III. Concrete relaxation of MAP problem

A. Reparametrization

In this publication, the idea is to use the concrete distribution in order to relax the MAP problem (2c) to

$$\hat{\mathbf{x}} = \arg\min_{\tilde{\mathbf{x}} \in \mathbb{R}^{N_\mathrm{T}\times 1}} -\ln p(\mathbf{y}|\tilde{\mathbf{x}}) - \ln p(\tilde{\mathbf{x}}) . \tag{7}$$

The reparametrization property makes it possible to express each $\tilde{x}_n$ in $\tilde{\mathbf{x}}$ by a vector $\mathbf{g}_n$ of i.i.d. Gumbel RVs instead:

$$\tilde{\mathbf{x}}(\mathbf{G}) = \begin{bmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_{N_\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{z}}_1^T \\ \vdots \\ \tilde{\mathbf{z}}_{N_\mathrm{T}}^T \end{bmatrix} \mathbf{m} = \begin{bmatrix} f\left(\mathbf{g}_1\right)^T \\ \vdots \\ f\left(\mathbf{g}_{N_\mathrm{T}}\right)^T \end{bmatrix} \mathbf{m} \tag{8}$$

$$\text{with} \quad \mathbf{G} = \begin{bmatrix} \mathbf{g}_1 & \cdots & \mathbf{g}_{N_\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{M\times N_\mathrm{T}} . \tag{9}$$

Now, we reformulate the relaxed MAP problem (7) and optimize w.r.t. matrix $\mathbf{G}$. This means, we replace $p(\mathbf{y}|\tilde{\mathbf{x}})$

by $p(\mathbf{y}|\mathbf{G})$ and introduce the Gumbel distribution $p(g_{kn}) = \exp\left(-g_{kn} - \exp\left(-g_{kn}\right)\right)$ as the new prior distribution:

$$\begin{aligned}
\hat{\mathbf{G}} &= \arg\min_{\mathbf{G}} -\ln p(\mathbf{y}|\mathbf{G}) - \ln p(\mathbf{G}) && (10)\\
&= \arg\min_{\mathbf{G}} -\ln p(\mathbf{y}|\mathbf{G}) - \sum_{n=1}^{N_{\mathrm{T}}}\sum_{k=1}^{M} \ln p(g_{kn}) && (11)\\
&= \arg\min_{\mathbf{G}} \frac{1}{2\sigma_{\mathrm{n}}^2}(\mathbf{y}-\mathbf{H}\tilde{\mathbf{x}}(\mathbf{G}))^T(\mathbf{y}-\mathbf{H}\tilde{\mathbf{x}}(\mathbf{G}))\\
&\quad + \mathbf{1}^T\mathbf{G}\mathbf{1} + \mathbf{1}^T e^{-\mathbf{G}}\mathbf{1} && (12)\\
&= \arg\min_{\mathbf{G}} L(\mathbf{G})\,. && (13)
\end{aligned}$$

However, owing to the softmax and exponential terms in $L(\mathbf{G})$, the objective (13) has no analytical solution. Furthermore, it describes a non-convex objective function which is illustrated in Fig. 2 for the binary case $M = 2$. This results from log-convexity of the concrete distribution for $\tau \leq (M-1)^{-1}$ [5]. The conditional pdf $p(\mathbf{y}|\tilde{\mathbf{x}})$ is log-concave and the prior pdf $p(\tilde{\mathbf{x}})$ log-convex, so the negative log joint distribution forms a non-convex objective function (13).

B. Gradient Descent Algorithm

In order to solve (13) with relatively low complexity, we employ the gradient descent approach. It tries to approach the minimum iteratively by taking gradient descent steps until the necessary condition

$$\frac{\partial L(\mathbf{G})}{\partial \mathbf{G}} = \mathbf{0} \tag{14}$$

is fulfilled. We point out that convergence to the global solution depends heavily on the starting point initialization since the objective function is non-convex. Without any prior information about the RV $\mathbf{x}$, the optimal starting point is $\mathbf{G}^{(0)} = \mathbf{0}$ since $f(\mathbf{0}) = \boldsymbol{\alpha}$ with $\tau = 1$. After some tensor/matrix calculus and by noting that every $\tilde{x}_n$ only depends on one $\mathbf{g}_n$, the gradient descent step for (13) in iteration j is:

$$\begin{aligned}
\mathbf{G}^{(j+1)} =& \mathbf{G}^{(j)} - \delta^{(j)} \cdot \left.\frac{\partial L(\mathbf{G})}{\partial \mathbf{G}}\right|_{\mathbf{G}=\mathbf{G}^{(j)}} && (15)\\
\frac{\partial L(\mathbf{G})}{\partial \mathbf{G}} =& \frac{1}{\sigma_{\mathrm{n}}^2} \cdot \begin{bmatrix} \frac{\partial \tilde{x}_1(\mathbf{g}_1)}{\partial \mathbf{g}_1} & \cdots & \frac{\partial \tilde{x}_{N_{\mathrm{T}}}(\mathbf{g}_{N_{\mathrm{T}}})}{\partial \mathbf{g}_{N_{\mathrm{T}}}} \end{bmatrix}\\
& \cdot \operatorname{diag}\left\{\mathbf{H}^T\mathbf{H}\tilde{\mathbf{x}}(\mathbf{G}) - \mathbf{H}^T\mathbf{y}\right\}\\
& + \mathbf{1}\cdot\mathbf{1}^T - e^{-\mathbf{G}} && (16)\\
\frac{\partial \tilde{x}_n(\mathbf{g}_n)}{\partial \mathbf{g}_n} =& \frac{1}{\tau^{(j)}} \cdot \left[\operatorname{diag}\left\{f\left(\mathbf{g}_n\right)\right\} - f\left(\mathbf{g}_n\right)\cdot f\left(\mathbf{g}_n\right)^T\right] \cdot \mathbf{m}\,. && (17)
\end{aligned}$$

Here, the operator diag $\{\mathbf{a}\}$ creates a diagonal matrix with the vector $\mathbf{a}$ on its main diagonal. The step-size $\delta^{(j)}$ can be chosen adaptively in every iteration j just as the hyperparameter $\tau^{(j)}$. For example, we can follow a heuristic schedule like in simulated annealing: We start with a large $\tau^{(j)}$ and decrease until we approach the true prior pdf for $\tau^{(j)} \to 0$. In the following, we denote our approach as Concrete MAP Detection (CMD). It is a generic approach applicable in any nonlinear differentiable probabilistic model. Furthermore, only elementwise nonlinearities and matrix vector multiplications are present. In particular, the matrix vector multiplications in $\mathbf{H}^T\mathbf{H}\tilde{\mathbf{x}}$ and operations in $f\left(\mathbf{g}_n\right) f\left(\mathbf{g}_n\right)^T \mathbf{m}$ have the largest impact on complexity of CMD. This implies a linear iterative complexity $\mathcal{O}(N_{\mathrm{T}} \cdot (2N_{\mathrm{R}} + 4M))$, i.e., CMD scales linearly with the input and output dimension as well as the number of classes. Hence, CMD exhibits a much lower complexity compared to that of MAP detection (2c) of $\mathcal{O}(M^{N_{\mathrm{T}}} \cdot N_{\mathrm{T}}N_{\mathrm{R}})$ growing exponentially with the number of elements in $\mathbf{x}$.

C. Special Case: Binary Random Variables

In order to interprete the algorithm, we now focus on the special case of binary RVs. We have $M = 2$ classes and hence only one degree of freedom in the softmax function (5):

$$\begin{aligned}
\tilde{x}_n(\mathbf{g}_n) = \tilde{\mathbf{z}}_n^T\mathbf{m} &= \begin{bmatrix} \tilde{z}_{1n} & \tilde{z}_{2n} \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix}\\
&= \begin{bmatrix} \tilde{z}_{1n} & 1-\tilde{z}_{1n} \end{bmatrix} \cdot \begin{bmatrix} -1 \\ 1 \end{bmatrix}\\
&= -2\tilde{z}_{1n} + 1\,. && (18)
\end{aligned}$$

After rewriting

$$\begin{aligned}
\tilde{z}_{1n} = f\left([g_{1n}, g_{2n}]^T\right) &= \frac{e^{\frac{\ln\alpha_1 + g_{1n}}{\tau}}}{e^{\frac{\ln\alpha_1 + g_{1n}}{\tau}} + e^{\frac{\ln\alpha_2 + g_{2n}}{\tau}}}\\
&= \frac{1}{1 + e^{\frac{\ln\alpha_2 - \ln\alpha_1 + g_{2n} - g_{1n}}{\tau}}}\,, && (19)
\end{aligned}$$

we notice that the difference of two i.i.d. Gumbel RVs $s_n = g_{2n} - g_{1n}$ is distributed according to the logistic distribution $p(s) = \exp(-s)/(1+\exp(-s))^2$. By transforming the two Gumbel RVs g_{2n} and g_{1n} into one stochastic variable s_n and making use of $\alpha = \alpha_1 = 1 - \alpha_2$, we have

$$\tilde{z}_{1n} = \frac{1}{1 + e^{\frac{\ln(1/\alpha - 1) + s_n}{\tau}}}\,. \tag{20}$$

Finally, we combine (18) and (20) to arrive at

$$\tilde{\mathbf{x}}(\mathbf{s}) = \tanh\left(\frac{\ln\left(1/\alpha - 1\right) + \mathbf{s}}{2\tau}\right)\,. \tag{21}$$

Now, we reparametrize the objective function for binary RVs in terms of logistic RVs $\mathbf{s} \in \mathbb{R}^{N_{\mathrm{T}}\times 1}$ with the new a-priori pdf $p(\mathbf{s})$:

$$\begin{aligned}
L(\mathbf{s}) =& -\ln p(\mathbf{y}|\mathbf{s}) - \sum_{n=1}^{N_{\mathrm{T}}} \ln p(s_n) && (22)\\
=& \frac{1}{2\sigma_{\mathrm{n}}^2}(\mathbf{y}-\mathbf{H}\tilde{\mathbf{x}}(\mathbf{s}))^T(\mathbf{y}-\mathbf{H}\tilde{\mathbf{x}}(\mathbf{s}))\\
& + \mathbf{1}^T\mathbf{s} + 2\cdot\mathbf{1}^T\ln\left(1 + e^{-\mathbf{s}}\right)\,. && (23)
\end{aligned}$$

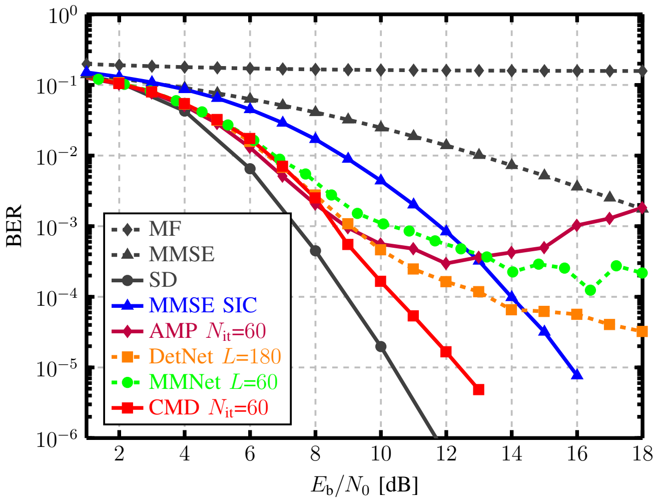

Fig. 3. BER curves of several detection methods in a 30×30 MIMO system with QPSK modulation. Effective system dimension is 60×60.

The prior pdf acts like a regularization contributing the second term. Analogously, we derive the gradient descent step of binary CMD:

$$\mathbf{s}^{(j+1)} = \mathbf{s}^{(j)} - \delta^{(j)} \cdot \left. \frac{\partial L(\mathbf{s})}{\partial \mathbf{s}} \right|_{\mathbf{s}=\mathbf{s}^{(j)}} \tag{24}$$

$$\frac{\partial L(\mathbf{s})}{\partial \mathbf{s}} = \frac{1}{\sigma_\mathrm{n}^2} \cdot \frac{\partial \tilde{\mathbf{x}}(\mathbf{s})}{\partial \mathbf{s}} \cdot \left[\mathbf{H}^T \mathbf{H} \tilde{\mathbf{x}}(\mathbf{s}) - \mathbf{H}^T \mathbf{y} \right] + \tanh\left(\frac{\mathbf{s}}{2} \right) \tag{25}$$

$$\frac{\partial \tilde{\mathbf{x}}(\mathbf{s})}{\partial \mathbf{s}} = \frac{1}{2\tau^{(j)}} \cdot \mathrm{diag}\left\{ 1 - \tilde{\mathbf{x}}^2(\mathbf{s}) \right\} . \tag{26}$$

Again, we notice that the iterative complexity of $\mathcal{O}(2N_\mathrm{T}N_\mathrm{R})$ of binary CMD is much lower than the MAP detection complexity of $\mathcal{O}(2^{N_\mathrm{T}} \cdot N_\mathrm{T}N_\mathrm{R})$.

IV. Numerical Results

In order to evaluate the performance of the proposed approach, we present numerical simulation results for application in MIMO systems with N_T transmit and N_R receive antennas. In this paper, we restrict to QPSK transmissions with Gray encoding and transform the complex-valued into the equivalent real-valued system model (1) so that we have $\mathbf{x} \in \mathcal{M}^{2N_\mathrm{T} \times 1}$ and $x_n \in \{\pm 1\}$. The model allows to compare to Deep Neural Network (DNN) based approaches for MAP detection in MIMO transmissions [7], [8], [9]. Hence, we test Concrete MAP Detection (CMD) for the binary case. The step size $\delta^{(j)}$ and softmax temperature $\tau^{(j)}$ are chosen to minimize the cross-entropy of $\mathbf{x}$. That means, we learn $\delta^{(j)}$ and $\tau^{(j)}$ using data and applying stochastic gradient descent methods. As an example, we assume the number of iterations to be $N_\mathrm{it} = 2N_\mathrm{T}$. Furthermore, we compare CMD to several State of the Art (SotA) approaches for MIMO detection. Fig. 3 shows the results in a large symmetric 30×30 MIMO system in terms of Bit Error Rates (BER) as a function of E_b/N_0. For QPSK, $E_\mathrm{b}/N_0 = 10\log_{10}(1/\sigma_\mathrm{n}^2) - 3\mathrm{dB}$.

Apparently, linear detectors perform bad in this setup: The curve of the Matched Filter (MF) remains almost constant at BER $\approx 20\%$. The Zero Forcer performs even worse and is hence not shown. At least, the MMSE estimator shows acceptable behavior but is still separated by a 7 dB gap from the optimal performance of the Sphere Detector (SD).

Nonlinear SotA detectors show good performance. For example, the BER of MMSE Successive Interference Cancellation with sorted QR decomposition and post sorting (MMSE SIC) from [10] decreases much faster. However, a 5 dB gap still remains as the low-complexity approach is only well-suited for small system dimensions. In contrast, Approximate Message Passing (AMP) is also of low complexity and optimal for large system dimensions [3]. This becomes evident at small $E_\mathrm{b}/N_0 < 8$ dB, where the BER curve is 1 dB close to that of the SD. At high $E_\mathrm{b}/N_0 > 8$ dB, the AMP runs into an error floor since the message statistics are not Gaussian anymore in finite small-scale MIMO systems.

Notably, our approach CMD offers the best performance of considered detection methods. It performs only slightly worse than the AMP at low SNR and does not run into an error floor in the simulated SNR range. The BER curve for $N_\mathrm{it} = 60$ decreases very fast and the performance loss compared to SD only amounts to 1-2 dB at high SNR. Additionally, the complexity is comparable to that of the AMP due to the similar algorithmic structure. An exact complexity comparison is left for future work.

We conclude our investigation by a comparison with results of latest research. To the best of our knowledge, DetNet [7], [8] is one of the first approaches transferring the results of Deep Learning research into communication systems. It is a very complex DNN architecture with a large number of parameters based on a projected gradient descent. The BER curve of DetNet with $L = 180$ layers is similar at low SNR but decreases more slowly than that of CMD at high SNR. Inspired from AMP, the authors from [9] propose a new DNN-like network MMNet with 2 parameters per layer like CMD. Designed for massive MIMO systems similar to DetNet, MMNet with the same number of iterations $L = 60$ fails to beat CMD. Surprisingly, MMNet works even worse than its inspiration, the AMP, at low SNR. In contrast to CMD, both approaches run into an error floor early.

For the sake of completeness, we also show results for a smaller 10×10 MIMO system in Fig. 4. Now, all soft nonlinear approaches run into an error floor at lower SNR. Thus, we conjecture that they share the same suboptimality at finite system dimensions which requires further research. However, CMD offers still the best overall performance and is even better than MMSE SIC for $E_\mathrm{b}/N_0 < 10$ dB.

V. Conclusion

In this paper, we presented Concrete MAP Detection (CMD). By means of the machine learning inspired continuous relaxation of discrete random variables, the concrete distribution, we relaxed the discrete MAP problem. This offers many favorable properties such as a differentiable objective function

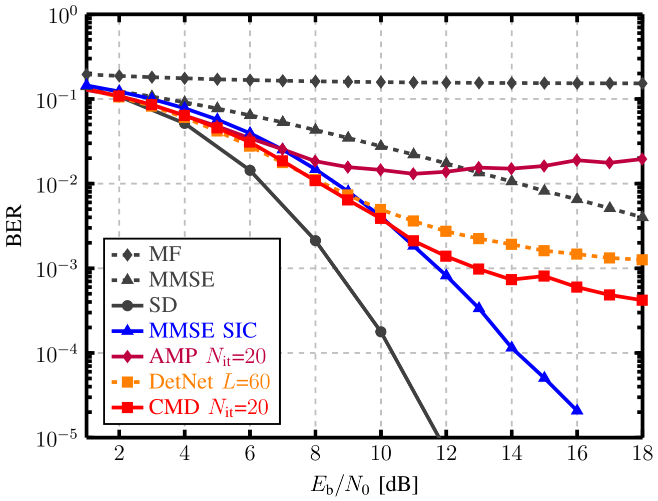

Fig. 4. BER curves of several detection methods in a 10×10 MIMO system with QPSK modulation. Effective system dimension is 20×20.

enabling a gradient descent based optimization. CMD beats the performance of the considered SotA approaches in large MIMO systems. In contrast to recent DNN based approaches, it offers a low iterative complexity comparable to that of the AMP as well. We leave extensions of CMD, e.g., to massive MIMO systems, and a detailed complexity analysis for future research.

References

[1] O. Simeone, “A Very Brief Introduction to Machine Learning with Applications to Communication Systems,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 648–664, Dec. 2018.

[2] ——, “A Brief Introduction to Machine Learning for Engineers,” *arXiv preprint arXiv:1611.01144*, Sep. 2017. [Online]. Available: https://arxiv.org/abs/1709.02840

[3] C. Jeon, R. Ghods, A. Maleki, and C. Studer, “Optimal Data Detection in Large MIMO,” *arXiv preprint arXiv:1811.01917*, Nov. 2018. [Online]. Available: https://arxiv.org/abs/1811.01917

[4] D. Wübben, “Effiziente Detektionsverfahren für Multilayer-MIMO-Systeme,” Ph.D.-Thesis, University of Bremen, Germany, Feb. 2006.

[5] C. J. Maddison, A. Mnih, and Y. W. Teh, “The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables,” *arXiv preprint arXiv:1611.00712*, Nov. 2016. [Online]. Available: https://arxiv.org/abs/1611.00712

[6] E. Jang, S. Gu, and B. Poole, “Categorical Reparameterization with Gumbel-Softmax,” *arXiv preprint arXiv:1611.01144*, Nov. 2016. [Online]. Available: https://arxiv.org/abs/1611.01144

[7] N. Samuel, T. Diskin, and A. Wiesel, “Deep MIMO Detection,” in *2017 IEEE 18th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Sapporo, Japan, Jul. 2017, pp. 1–5.

[8] ——, “Learning to Detect,” *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2554–2564, May 2019.

[9] M. Khani, M. Alizadeh, J. Hoydis, and P. Fleming, “Adaptive Neural Signal Detection for Massive MIMO,” *arXiv preprint arXiv:1906.04610*, Jun. 2019. [Online]. Available: https://arxiv.org/abs/1906.04610

[10] D. Wübben, R. Böhnke, V. Kühn, and K.-D. Kammeyer, “MMSE Extension of V-BLAST based on Sorted QR Decomposition,” in *2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No. 03CH37484)*, vol. 1, Orlando, USA, Oct. 2003, pp. 508–512.