Méthode du point proximal: principe et applications aux algorithmes itératifs

Ziad NAJA[1], Florence ALBERGE[1], Pierre DUHAMEL[2]

[1]Univ Paris-Sud 11

[2]CNRS
Laboratoire des signaux et systèmes (L2S)
Supelec, 3 rue Joliot-Curie 91192 Gif-sur-Yvette cedex (France)
Ziad.Naja@lss.supelec.fr, Florence.Alberge@lss.supelec.fr,
Pierre.Duhamel@lss.supelec.fr

Résumé – Cet article est basé sur l'algorithme du point proximal. Nous étudions deux algorithmes itératifs: l'algorithme de Blahut-Arimoto communément utilisé pour le calcul de la capacité des canaux discrets sans mémoire puis le décodage itératif pour les modulations codées à bits entrelacés. Dans les deux cas, il s'agit d'algorithmes itératifs pour lesquels les méthodes de type point proximal conduisent à une nouvelle interprétation et ouvrent la voie à des améliorations en terme de vitesse de convergence notamment.

Abstract – This paper recalls the proximal point method. We study two iterative algorithms: the Blahut-Arimoto algorithm for computing the capacity of arbitrary discrete memoryless channels, as an example of an iterative algorithm working with probability density estimates and the iterative decoding of the Bit Interleaved Coded Modulation (BICM-ID). For these iterative algorithms, we apply the proximal point method which allows new interpretations with improved convergence rate.

1 Introduction

Cet article s'intéresse à deux algorithmes itératifs classiques : l'algorithme de Blahut-Arimoto [1, 2] pour le calcul de la capacité d'un canal discret sans mémoire et le décodage itératif des modulations codées à bits entrelacés (BICM-ID) [3]. Bien que ces méthodes soient radicalement différentes à la fois par l'application visée et aussi par le processus itératif mis en jeu, elles ont pour point commun de présenter des connections avec une méthode d'optimisation bien connue, la méthode du point proximal [4].
En 1972, R. Blahut et S. Arimoto [1, 2] ont montré comment calculer numériquement la capacité des canaux sans mémoire avec des entrées et des sorties à alphabets finis. Depuis, plusieurs extensions ont été proposées citons notemment [5] qui a étendu l'algorithme de Blahut-Arimoto aux canaux avec mémoire et entrées à alphabets finis et [6] qui a considéré des canaux sans mémoire avec des entrées et/ou des sorties continues.
En parallèle, d'autres travaux se sont concentrés sur l'interprétation géométrique de l'algorithme de Blahut-Arimoto [7]. En se basant sur cette dernière approche, Matz [8] a proposé une version modifiée de cet algorithme qui converge plus vite que l'algorithme standard.
L'algorithme proposé par Matz est basé sur une approximation d'un algorithme de point proximal. Nous proposons donc dans ce qui suit une vrai reformulation point proximal avec une vitesse de convergence plus grande comparée à celle de l'algorithme classique de Blahut-Arimoto ainsi qu'à celle de l'approche dans [8].
D'autre part, les modulations codées à bits entrelacés (BICM) ont été d'abord proposés par Zehavi [9] pour améliorer la performance des modulations codées en treillis dans le cas des canaux de Rayleigh à évanouissement. Le décodage itératif [10] utilisé pour les BICM a une structure similaire à celle d'un turbo décodeur série. Bien que très performant, le décodage itératif n'a pas été à l'origine introduit comme solution d'un problème d'optimisation, ce qui rend difficile l'analyse de sa convergence.
Cet article va donc mettre en évidence le lien existant entre ces deux algorithmes itératifs et montrer comment cela conduit à des améliorations substantielles tout en révélant le lien existant entre le décodage itératif et les techniques classiques d'optimisation.

2 Algorithme du point proximal

L'algorithme du point proximal, dans sa version d'origine, est caractérisé par le processus itératif [11] :

$$\theta^{(k+1)} = \arg\max_{\theta}\{\xi(\theta) - \beta_k\|\theta - \theta^{(k)}\|^2\} \qquad (1)$$

dans lequel $\xi(\theta)$ est la fonction de coût qui croît au fil des itérations et $\|\theta - \theta^{(k)}\|^2$ est un terme de pénalité qui assure que la nouvelle valeur du paramètre reste dans le voisinage de la valeur obtenue à l'itération précédente. $\{\beta_k\}_{k\geq 0}$ est une séquence

de paramètres positifs. lorsque la séquence β_k converge vers zéro à l'infini, alors la méthode présente une convergence superlinéaire [12]. L'algorithme du point proximal peut être généralisé selon :

$$\theta^{(k+1)} = \arg\max_{\theta}\{\xi(\theta) - \beta_k f(\theta, \theta^{(k)})\}$$

où $f(\theta, \theta^{(k)})$ est toujours non négative et $f(\theta, \theta^{(k)}) = 0$ si et seulement si $\theta = \theta^{(k)}$. Dans la suite, nous utiliserons cette formulation en considérant pour f soit la divergence de Kullback soit la divergence de Fermi-Dirac. Nous rappelons maintenant leurs définitions.

La distance de Kullback-Leibler (KLD) est définie pour deux distributions de probabilité $p = \{p(x), x \in \boldsymbol{X}\}$ et $q = \{q(x), x \in \boldsymbol{X}\}$ d'une variable aléatoire discrète $\mathbf{X}$ prenant ses valeurs $\mathbf{x}$ dans un ensemble discret $\boldsymbol{X}$ par :

$$D(p||q) = \sum_{\mathbf{x}\in\boldsymbol{X}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}$$

La distance de Kullback (appelée aussi entropie relative) a deux propriétés importantes : $D(p||q)$ est toujours non-négative, et $D(p||q)$ est nulle si et seulement si $p = q$. Cependant, ce n'est pas une "vraie" distance puisqu'elle n'est pas symétrique ($D(p||q) \neq D(q||p)$) et ne satisfait pas en général l'inégalité triangulaire.

La divergence de Fermi-Dirac est la divergence de Kullback-Leibler appliquée à des probabilités sur des évènements n'ayant que deux issues, elle est définie pour deux distributions de probabilité $r_i = P_R(x_i = 1)$ et $s_i = P_S(x_i = 1)$ définies dans l'ensemble $\mathbf{X} = (x_1, \ldots, x_n)$ avec $x_i \in \{0, 1\}$ de la manière suivante :

$D_{FD}(\mathbf{r}, \mathbf{s}) = \sum_{i=1}^{n} r_i \log\left(\frac{r_i}{s_i}\right) + \sum_{i=1}^{n}(1 - r_i) \log\left(\frac{1-r_i}{1-s_i}\right)$

La divergence de Fermi-Dirac présente les deux mêmes propriétés que la distance de Kullback : $D_{FD}(\mathbf{r}, \mathbf{s})$ est toujours non négative et $D_{FD}(\mathbf{r}, \mathbf{s}) = 0$ si et seulement si $\mathbf{r} = \mathbf{s}$. La divergence de Fermi-Dirac n'est pas symétrique.

3 Méthode de point proximal pour les algorithmes itératifs

3.1 Algorithme de Blahut-Arimoto [1] et interprétation point proximal

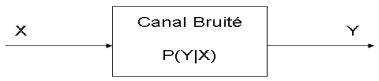

FIG. 1 – Canal.

Considérons un canal discret sans mémoire avec pour entrée X prenant ses valeurs dans l'ensemble $\{x_0, \ldots, x_M\}$ et en sortie Y prenant ses valeurs dans l'ensemble $\{y_0, \ldots, y_N\}$. Ce canal est défini par sa matrice de transition Q telle que $[Q]_{ij}$ = $Q_{i|j} = Pr(Y = y_i|X = x_j)$.

Nous définissons aussi $p_j = Pr(X = x_j)$ et $q_i = Pr(Y = y_i)$. L'information mutuelle est donnée par : $I(X,Y) = I(p,Q) = \sum_{j=0}^{M}\sum_{i=0}^{N} p_j Q_{i|j} \log \frac{Q_{i|j}}{q_i} = \sum_{j=0}^{M} p_j D(Q_j||q)$ et la capacité du canal par :

$$C = \max_{p} I(p, Q)$$

En résolvant ce problème de maximisation et en prenant en compte la condition de normalisation, nous obtenons le processus itératif :

$$p^{(k+1)}(x) = \frac{p^{(k)}(x)\exp(D_x^k)}{\sum_x^M p^{(k)}(x)\exp(D_x^k)} \quad (2)$$

avec $D_x^k = D(p(Y = y|X = x)||p(Y = y^{(k)}))$. C'est l'algorithme de Blahut-Arimoto. On peut montrer sans difficulté que cet algorithme est équivalent à :

$$p^{(k+1)}(x) = \arg\max_{p}\{I^{(k)}(p(x)) - D(p(x)||p^{(k)}(x))\} \quad (3)$$

où $I^{(k)}(p(x)) = \mathbb{E}_{p(x)}\{D_x^k\}$. Cet algorithme n'est pas un algorithme du point proximal puisque la fonction de coût $I^{(k)}(p(x))$ dépend des itérations. Il est toutefois possible d'exprimer l'information mutuelle comme suit :

$$I(p(x)) = I^{(k)}(p(x)) - D(q(y)||q^{(k)}(y)) \quad (4)$$

En introduisant (4) dans (3), nous obtenons :

$$p^{(k+1)}(x) = \arg\max_{p}\{I(p(x)) - (D(p(x)||p^{(k)}(x)) - D(q(y)||q^{(k)}(y)))\}$$

D'après l'inégalité de Jensen, nous pouvons montrer que le terme de pénalité

$$D(p(x)||p^{(k)}(x)) - D(q(y)||q^{(k)}(y)) = \mathbb{E}_{p(x,y)}[\log \frac{p(x)\sum_{\tilde{x}} p(y|\tilde{x})p^{(k)}(\tilde{x})}{p^{(k)}(x)\sum_{\tilde{x}} p(y|\tilde{x})p(\tilde{x})}]$$

est toujours positif et qu'il est nul si et seulement si $p(x) = p^{(k)}(x)$ et $q(y) = q^{(k)}(y)$.

Le processus itératif devient alors :

$$p^{(k+1)}(x) = \arg\max_{p(x)}\{I(p(x)) - \beta_k\{D(p(x)||p^{(k)}(x)) - D(q(y)||q^{(k)}(y))\}\}$$

A chaque itération, l'expression de $p^{(k+1)}(x)$ est la même que dans (2). L'algorithme de Blahut-Arimoto s'interprète donc comme un algorithme du point proximal dans lequel le paramètre β_k est constant et égal à 1.

L'approche intuitive de Matz [8] consiste à remplacer la distribution de probabilité $q(y)$ dans le terme de droite de l'équation précédente par la même distribution $q^{(k)}(y)$ calculée à l'itération précédente.

Nous allons maintenant utiliser le degré de liberté supplémentaire amené par β_k pour augmenter la vitesse de convergence. Nous choisissons β_k comme suit :

$$\max_{\beta_k} \beta_k(D(p^{(k+1)}(x)||p^{(k)}(x)) - D(q^{(k+1)}(y)||q^{(k)}(y)))$$

dans lequel $p^{(k+1)}(x)$ et $q^{(k+1)}(y)$ dépendent de β_k. Cela garantie que $I(p^{(k+1)}(x)) - I(p^{(k)}(x))$ est maximale à chaque itération. Pour résoudre ce problème de maximisation, nous avons utilisé la méthode de gradient conjugué qui donne la valeur de β_k la plus convenable en comparaison avec l'approche proposée par Matz.

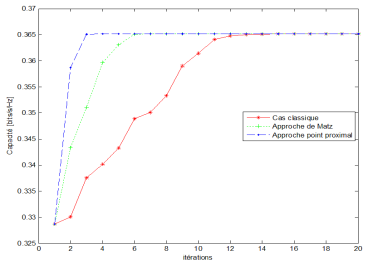

FIG. 2 – Canal discret binaire symétrique.

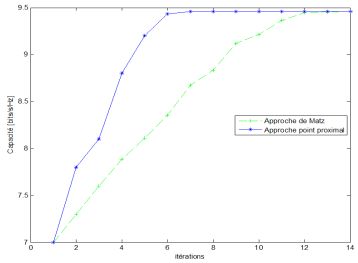

FIG. 3 – Canal Gaussian Bernouilli-Gaussian ayant comme paramètres $(p = 0.3, \sigma_b = 0.01, \sigma_g = 1)$.

3.1.1 Simulation

Nous testons les 3 algorithmes itératifs sur un canal discret binaire symétrique défini par sa matrice de transition :

$$Q = \left\{ \begin{array}{ccc} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.2 & 0.7 \end{array} \right\}$$

Les résultats (fig.2) montrent que la capacité du canal est atteinte après 20 itérations dans le cas classique, 7 itérations dans l'approche de Matz et 4 itérations dans notre cas (avec une précision de 10^{-11}).

Nous comparons ensuite notre algorithme et celui de Matz dans le cas d'un canal Gaussian Bernouilli-Gaussian dans le but de former une matrice Q avec de grandes dimensions. Un tel canal est défini par : $y_k = x_k + b_k + \gamma_k$ où

- $b \sim \mathcal{N}(0, \sigma_b^2)$
- $\gamma_k = e_k g_k$ avec e : séquence de Bernouilli(p)
- $g \sim \mathcal{N}(0, \sigma_g^2)$ avec $\sigma_b^2 \ll \sigma_g^2$

d'où

$$y_k = x_k + n_k$$

avec

$$p(n_k) = (1-p)\mathcal{N}(0, \sigma_b^2) + p\mathcal{N}(0, \sigma_b^2 + \sigma_g^2)$$

La sortie y_k a été discrétisée sur 40 valeurs, et l'entrée x_k sur 10 valeurs. Les résultats sont reportés sur la figure 3. Nous observons encore un gain conséquent grâce à notre approche.

3.2 Outils de base

Nous introduisons tout d'abord quelques notations. Soit $\mathbf{B}_i \in \{0,1\}^N$ la représentation binaire d'un entier $i, 0 \leq i \leq 2^{N-1}$. $\mathbf{B} = (\mathbf{B}_0, \mathbf{B}_1, ..., \mathbf{B}_{2^N-1})^T$ de dimension $2^N \times N$ est la matrice de la représentation binaire de tous les mots de longueur N. Soit η la fonction densité de probabilité de la variable $\chi = \mathbf{B}_i$. On a donc

$$\eta = (\mathbf{Pr}[\chi = \mathbf{B_0}], \mathbf{Pr}[\chi = \mathbf{B_1}], ..., \mathbf{Pr}[\chi = \mathbf{B_{2^N-1}}])^{\mathbf{T}}$$

Etant donné une fonction densité de probabilité η, ses coordonnées logarithmiques sont le vecteur θ dont le i^{eme} élément est donné par $\theta_i = \ln(Pr[\chi = \mathbf{B_i}]) - \ln(Pr[\chi = \mathbf{B_0}])$. Nous définissons aussi λ le vecteur des ratio dont l'élément j est défini par $\lambda_j = log(\frac{Pr[\chi_j=1]}{Pr[\chi_j=0]})$ où χ_j est le j^{eme} bit du mot binaire χ et $\lambda \in \mathbb{R}^N$. Pour des densités séparables, c'est à dire qui sont égales au produit des marginales, les coordonnées logarithmiques prennent la forme $\theta = \mathbf{B}\lambda$ [13].

3.2.1 Décodage itératif des modulations codées à bits entrelacés [3]

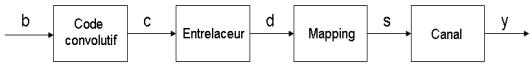

FIG. 4 – Codeur des modulations codées à bits entrelacés.

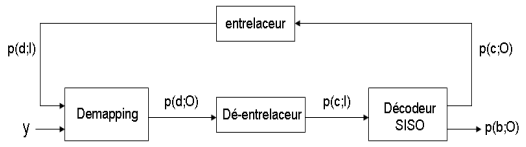

FIG. 5 – Décodeur itératif des modulations codées à bits entrelacés.

Le décodage itératif pour les modulations codées à bits entrelacées est constitué de deux blocs chacun ayant pour tâche d'évaluer des probabilités a posteriori. Le premier bloc (demapping) contient les informations concernant le mapping et le canal au travers de la loi de probabilité $p(\mathbf{y}|\mathbf{s})$ où $\mathbf{y}$ est le vecteur reçu et $\mathbf{s}$ un vecteur de symbole. Ce bloc reçoit un a priori (aussi appelé extrinsèque) qui lui est fourni par l'autre bloc. Il est donc en mesure de fournir des probabilités à posterori que nous noterons $p_{\mathbf{B}\lambda_1+\theta_m}$ où $(\lambda_1)_{km+i} = \ln\left(\frac{p(d_{km+i}=1;I)}{p(d_{km+i}=0;I)}\right)$ est le vecteur contenant les log-ratio de la probabilité a priori [13]. Le vecteur θ_m est le vecteur de coordonnées logarithmiques obtenu à partir de $p(\mathbf{y}|\mathbf{s})$. Le second bloc contient les informations correspondant au codeur au travers de la fonction indicatrice du code. Ce second bloc fournit les probabilités a posteriori sur les bits $p_{\mathbf{B}\lambda_2+\theta_c}$ où λ_2 dépend de l'a priori à l'entrée du bloc et θ_c est le vecteur de coordonnées logarithmiques obtenu à partir de la fonction indicatrice du code [13]. Par ailleurs, l'a priori du bloc suivant est calculé en divisant la probabilité a posteriori du bloc précédent par l'a priori qu'il a reçu (propagation d'extrinsèques). Ce principe peut être résumé par le processus itératif :

$$\text{Trouver } \lambda_2^{(k+1)} \text{ telle que } p_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k+1)})} = p_{\mathbf{B}\lambda_1^{(k)}+\theta_m} \quad (5)$$

$$\text{Trouver } \lambda_1^{(k+1)} \text{ telle que } p_{\mathbf{B}(\lambda_1^{(k+1)}+\lambda_2^{(k+1)})} = p_{\mathbf{B}\lambda_2^{(k+1)}+\theta_c} \quad (6)$$

Ce processus itératif correspond à la résolution du problème de minimisation suivant :
Au niveau du demapping

$$\min_{\lambda_2} D_{FD}(p_{\mathbf{B}\lambda_1+\theta_m}, p_{\mathbf{B}(\lambda_1+\lambda_2)})$$

Au niveau du décodeur

$$\min_{\lambda_1} D_{FD}(p_{\mathbf{B}\lambda_2+\theta_c}, p_{\mathbf{B}(\lambda_1+\lambda_2)})$$

Une solution est satisfaisante si elle répond aux deux critères simultanément.
Cependant la minimisation de l'un de ces critères n'entraine pas forcément la diminution de l'autre critère à l'itération suivante. On peut donc craindre un comportement de l'algorithme. La méthode du point proximal permet de faire le lien entre les deux critères via le terme de pénalité qu'elle introduit. Nous obtenons alors un nouveau processus de minimisation :

$$\lambda_2^{(k+1)} = \min_{\lambda_2} J_{\theta_m}(\lambda_1, \lambda_2) = \min_{\lambda_2} D_{FD}(\mathbf{p}_{\mathbf{B}\lambda_1+\theta_m}, \mathbf{p}_{\mathbf{B}(\lambda_1+\lambda_2)}) + \mu_m D_{FD}(\mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k)})}, \mathbf{p}_{\mathbf{B}(\lambda_1+\lambda_2)})$$

$$\lambda_1^{(k+1)} = \min_{\lambda_1} J_{\theta_c}(\lambda_1, \lambda_2) = \min_{\lambda_1} D_{FD}(\mathbf{p}_{\mathbf{B}\lambda_2+\theta_c}, \mathbf{p}_{\mathbf{B}(\lambda_1+\lambda_2)}) + \mu_c D_{FD}(\mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k+1)})}, \mathbf{p}_{\mathbf{B}(\lambda_1+\lambda_2)})$$

Cela revient à trouver $\lambda_2^{(k+1)}$ telle que

$$\mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k+1)})} = \frac{\mathbf{p}_{\mathbf{B}\lambda_1^{(k)}+\theta_m} + \mu_m \mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k)})}}{1+\mu_m} \quad (7)$$

et $\lambda_1^{(k+1)}$ telle que

$$\mathbf{p}_{\mathbf{B}(\lambda_1^{(k+1)}+\lambda_2^{(k+1)})} = \frac{\mathbf{p}_{\mathbf{B}\lambda_2^{(k+1)}+\theta_c} + \mu_c \mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k+1)})}}{1+\mu_c} \quad (8)$$

A la convergence, on retrouve les mêmes points stationnaires que pour (5) et (6). Pour assurer la décroissance des fonctions de coût, nous choisissons μ_m et μ_c afin que
$J_{\theta_m}(\lambda_1^{(k)}, \lambda_2^{(k+1)}) \leq J_{\theta_c}(\lambda_1^{(k)}, \lambda_2^{(k)})$ et
$J_{\theta_c}(\lambda_1^{(k+1)}, \lambda_2^{(k+1)}) \leq J_{\theta_m}(\lambda_1^{(k+1)}, \lambda_2^{(k+1)})$.
La première inégalité est équivalente à
$J_{\theta_m}(\lambda_1^{(k)}, \lambda_2^{(k+1)}) \leq \frac{\mu_m}{1+\mu_m}(D_{FD}(\mathbf{p}_{\mathbf{B}\lambda_1^{(k)}+\theta_m}, \mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k)})}) + D_{FD}(\mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k)})}, \mathbf{p}_{\mathbf{B}\lambda_1^{(k)}+\theta_m}))$ car la distance de Fermi-Dirac est convexe par rapport à son deuxième paramètre. D'autre part
$D_{FD}(\mathbf{p}_{\mathbf{B}\lambda_2^{(k)}+\theta_c}, \mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k)})}) \leq J_{\theta_c}(\lambda_1^{(k)}, \lambda_2^{(k)})$
D'après ces deux relations, nous obtenons une borne supérieure pour μ_m :

$$\mu_m \leq \frac{D_{FD}(\mathbf{p}_{\mathbf{B}\lambda_2^{(k)}+\theta_c}, \mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k)})})}{\mathcal{D}_{FD} - D_{FD}(\mathbf{p}_{\mathbf{B}\lambda_2^{(k)}+\theta_c}, \mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k)})})}$$

où $\mathcal{D}_{FD}$ est une distance symétrique :
$\mathcal{D}_{FD} = D_{FD}(\mathbf{p}_{\mathbf{B}\lambda_1^{(k)}+\theta_m}, \mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k)})})$
$+ D_{FD}(\mathbf{p}_{\mathbf{B}(\lambda_1^{(k)}+\lambda_2^{(k)})}, \mathbf{p}_{\mathbf{B}\lambda_1^{(k)}+\theta_m})$
La borne supérieure pour μ_c peut être obtenue d'une façon similaire. En itérant (7) et (8) avec μ_c et μ_m choisis correctement nous obtenons un algorithme qui converge vers les même points que le décodage itératif classique (et qui a donc les mêmes performances en terme de taux d'erreur binaire) tout en diminuant au fil des itérations un critère désiré.

4 Conclusion

Dans cet article, nous avons d'abord mis en évidence l'algorithme itératif du point proximal. Nous avons ensuite présenté deux algorithmes itératifs différents à la fois par l'application visée et le processus itératif mis en jeu : l'algorithme itératif de Blahut-Arimoto et l'algorithme de décodage itératif des modulations codées à bits entrelacés. Une interprétation de ces deux algorithmes basée sur la méthode de point proximal a donc été proposée appuyée par des résultats de simulation.

Références

[1] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, pp. 14–20, 1972.

[2] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 18, pp. 460–473, 1972.

[3] G. Caire, G.Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inf. Theory*, vol. 4, pp. 927–946, May 1998.

[4] G. Vige, "Proximal-point algorithm for minimizing quadratic functions," INRIA,RR-2610, Tech. Rep., 1995.

[5] F. Dupuis, W. Yu, and F. Willems, "Arimoto-Blahut algorithms for computing channel capacity and rate-distortion with side-information," in *ISIT*, 2004.

[6] J. Dauwels, "On graphical models for communications and machine learning : Algorithms, bounds, and analog implementation," Ph.D. dissertation, May 2006.

[7] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedure," *Statistics and Decisions*, vol. supplement issue 1, pp. 205–237, 1984.

[8] G. Matz and P. Duhamel, "Information geometric formulation and interpretation of accelerated Blahut-Arimoto-Type algorithms," in *Proc. Information Theory Workshop*, 2004.

[9] E. Zehavi, "8-PSK trellis codes for a Rayleigh fading channel," *IEEE Trans. Commun.*, vol. 40, pp. 873–883, May 1992.

[10] X. Li, A. Chindapol, and J. Ritcey, "Bit interleaved coded modulation with iterative decoding and 8-PSK signaling," *IEEE trans Commun.*, vol. 50, pp. 1250–1257, Aug 2002.

[11] S. Chrétien and A. O. Hero, "Kullback Proximal Algorithms for Maximum LikelihoodEstimation," INRIA,RR-3756, Tech. Rep., Aug 1999.

[12] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM Journal on Control and Optimization*, vol. 14, pp. 877–898, 1976.

[13] J. M. Walsh, P. Regalia, and C. R. Johnson, "Turbo decoding as Iterative Constrained Maximum-Likelihood Sequence Detection," *IEEE Trans. Inf. Theory*, vol. 52, pp. 5426–5437, Dec. 2006.