Supplemental Details and Derivations For Differentiable Sparse Coding

David M. Bradley
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
dbradley@cs.cmu.edu

J. Andrew Bagnell
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
dbagnell@ri.cmu.edu

A Properties of the Unnormalized KL-divergence Prior

Many of the important properties of a regularization function can be understood by looking at the behavior of its partial derivatives, since at $\hat{w}$ the (sub)gradient of the regularization function cancels the gradient of the loss (1). In the case where B is the identity matrix and $D_L(x\|f(Bw)) = \frac{1}{2}\|x - Bw\|_2^2$, each $\hat{w}_i$ can be computed independently as $\hat{w}_i = x_i - \nabla_i D_P(\hat{w}\|p)$. Figure 1A plots $\hat{w}_i$ against x_i for various priors. The uniform prior has no effect and produces the line $\hat{w}_i = x_i$. An L_2 (gaussian) prior changes the slope of the line to $\hat{w}_i = (1-\lambda)x_i$, but does not change the sparsity of $\hat{w}$, as all elements are scaled equally. An L_1 prior however, does change the sparsity because its gradient is discontinuous at zero which forces $\hat{w}_i = 0$ while $|x_i| \leq \lambda$.

$$\nabla D_P(\hat{w}\|p) = -\frac{1}{\lambda}\nabla D_L\left(x\|f(B\hat{w})\right) \tag{1}$$

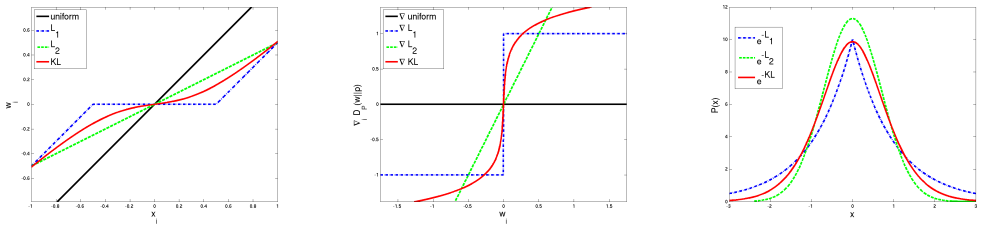

Figure 1: For the identity basis, the gradient of the log prior on $\hat{w}$ (B) determines an offset between $\hat{w}_i$ and x_i (A). The probability density functions obtained by exponentiating the L_1, L_2, and KL regularization functions are shown in (C). Note that KL regularization is shown on an expanded basis set that adds the negation of each basis vector to allow negative weights.

KL-divergence regularization does not allow negative weights, but expanding the basis by adding the negation of each basis vector $\widetilde{B} = [-B\ B]$ simulates the effects of negative weights with only positive ones. In this case the derivative of the KL-divergence prior corresponds to the arcsinh (w_i/p_i) function pictured in Figure 1B (proof given in Appendix B). arcsinh grows quickly for small weights, which causes sparsity similar to L_1. Crucially though, it grows slowly for large weights while still reaching ∞. This property causes $\hat{w}$ to be differentiable and stable to small changes in B and x, because it allows $\hat{w}$ to contain large weights while still ensuring that similar columns of B will have similar activations in $\hat{w}$. If B is not orthogonal, L_1 allows $\hat{w}$ to be discontinuous[1] with respect to X and B because its derivative is flat apart from $w = 0$.

[1]For an extreme, but illustrative, example consider the case where B contains two identical basis vectors. Then there is no longer even a unique $\hat{w}$.

B Relationship between KL-divergence and arcsinh

The MAP estimate $\widetilde{w}$ obtained for KL regularization on the basis $\widetilde{B} = [-B\ B]$ is related through the function $\hat{w} = \hat{w}^+ - \hat{w}^-$ to the MAP estimate $\hat{w}$ produced by using the regularization function $\sum_i w_i \text{arcsinh}\left(\frac{w_i}{2p_i}\right) - \sqrt{w_i^2 + 4p_i^2}$ (shown in figure A) on the basis B.

Define $\widetilde{B} = [-B\ B]$, and compute the MAP estimate $\widetilde{w}$ by minimizing (2).

$$\widetilde{w} = \arg\min_w D_L\left(x \| f(\widetilde{B}w)\right) + \lambda \text{KL}(\widetilde{w} \| p) \tag{2}$$

At the minimum $\widetilde{w}$, the gradients of both terms in (2) cancel each other (3).

$$-\frac{1}{\lambda}\frac{\partial}{\partial \widetilde{w}} D_L\left(x \| f(\widetilde{B}\widetilde{w})\right) = \frac{\partial}{\partial \widetilde{w}} \text{KL}(\widetilde{w} \| p) \tag{3}$$

Divide the elements of $\widetilde{w}$ into two groups $\widetilde{w} = \begin{bmatrix} \hat{w}^- \\ \hat{w}^+ \end{bmatrix}$, so that $\widetilde{B}\widetilde{w} = B(\hat{w}^+ - \hat{w}^-) = B\hat{w}$, where $\hat{w}$ is defined to be $\hat{w} = \hat{w}^+ - \hat{w}^-$. Notice that $\frac{\partial(\widetilde{B}\widetilde{w})}{\partial \widetilde{w}} = \left[-\frac{\partial B\hat{w}}{\partial w}\ \frac{\partial B\hat{w}}{\partial w}\right]$. Substituting $B\hat{w}$ into (3) produces (4).

$$-\frac{1}{\lambda}\begin{bmatrix} -\frac{\partial}{\partial \hat{w}} D_L\left(x \| f(B\hat{w})\right) \\ \frac{\partial}{\partial \hat{w}} D_L\left(x \| f(B\hat{w})\right) \end{bmatrix} = \begin{bmatrix} \log \frac{w^-}{p} \\ \log \frac{w^+}{p} \end{bmatrix} \tag{4}$$

Solve (4) for $\hat{w}_i^+$ and $\hat{w}_i^-$ (5).

$$\begin{aligned} \hat{w}^+ &= p e^{-\frac{1}{\lambda}\frac{\partial}{\partial \hat{w}} D_L(x \| f(B\hat{w}))} \\ \hat{w}^- &= p e^{\frac{1}{\lambda}\frac{\partial}{\partial \hat{w}} D_L(x \| f(B\hat{w}))} \end{aligned} \tag{5}$$

Now rewrite $\hat{w} = \hat{w}^+ - \hat{w}^-$ using (5) and the hyperbolic sin function $\sinh(x) = 1/2(e^x - e^{-x})$ (6).

$$\begin{aligned} \hat{w} &= p e^{-\frac{1}{\lambda}\frac{\partial}{\partial \hat{w}} D_L(x \| f(B\hat{w}))} - p e^{\frac{1}{\lambda}\frac{\partial}{\partial \hat{w}} D_L(x \| f(B\hat{w}))} \\ &= 2p \sinh\left(-\frac{1}{\lambda}\frac{\partial}{\partial \hat{w}} D_L\left(x \| f(B\hat{w})\right)\right) \end{aligned} \tag{6}$$

Rearranging (6) produces (7), which is the derivative of (8) at its MAP estimate. Hence $\hat{w}$ computed by KL-regularization is also the MAP estimate of (8).

$$-\frac{1}{\lambda}\frac{\partial}{\partial \hat{w}} D_L\left(x \| f(B\hat{w})\right) = \text{arcsinh}\left(\frac{\hat{w}}{2p}\right) \tag{7}$$

$$\hat{w} = \arg\min_w D_L\left(x \| f(Bw)\right) + \lambda \sum_i w_i \text{arcsinh}\left(\frac{w_i}{2p_i}\right) - \sqrt{w_i^2 + 4p_i^2} \tag{8}$$

C Derivation of $\frac{\partial \hat{w}}{\partial B}$

The KL-divergence prior we use does not have a closed-form solution for the MAP estimate of the coefficients, $\hat{w}$. However, the partial derivative $\frac{\partial \hat{w}}{\partial B}$ can still be computed using implicit differentiation because the gradient of the reconstruction loss equals the negative gradient of the regularization at $\hat{w}$ (1). This section derives $\frac{\partial \hat{w}}{\partial B}$ for general pairs of matching reconstruction loss $D_L\left(x \| r\right)$ and reconstruction transfer functions, $r = f(Bw)$, whose derivatives with respect to the coefficients assume a common form (9). Examples of such pairs include the linear output function with squared loss, and normalized exponential reconstruction with KL loss.

$$\frac{\partial D_L\left(x \| r\right)}{\partial w} = B^T\left(r - x\right) \tag{9}$$

The derivative of the gradient of the likelihood with respect to B_i^k is the vector equation (10), where $\vec{\mathrm{e}}_i$ is a unit vector whose ith element is 1.

$$\frac{\partial}{\partial B_i^k}\left(\frac{\partial D_L\left(x\|r\right)}{\partial w}\right) = B^T\left(\frac{\partial r}{\partial w}\frac{\partial w}{\partial B_i^k} + \frac{\partial r}{\partial B_i^k}\right) + \vec{\mathrm{e}}_i(r_k - x_k) \tag{10}$$

Similarly, the derivative of the gradient of the KL-divergence prior with respect to B_i^k is the product of a diagonal matrix and the column vector $\frac{\partial w}{\partial B_i^k}$ (11).

$$\frac{\partial}{\partial B_i^k}\left(\frac{\partial D_P(w\|p)}{\partial w}\right) = \mathrm{diag}\left(\frac{1}{w}\right)\frac{\partial w}{\partial B_i^k} \tag{11}$$

At the MAP estimate $\hat{w}$, (10) equals negative (11), and solving for $\frac{\partial \hat{w}}{\partial B_i^k}$ we get (12).

$$\begin{aligned}
\mathrm{diag}(\frac{-\lambda}{\hat{w}})\frac{\partial \hat{w}}{\partial B_i^k} &= B^T\left(\frac{\partial \hat{r}}{\partial \hat{w}}\frac{\partial \hat{w}}{\partial B_i^k} + \frac{\partial \hat{r}}{\partial B_i^k}\right) + \vec{\mathrm{e}}_i(\hat{r}_k - x_k) \\
-\left(B^T\frac{\partial \hat{r}}{\partial \hat{w}} + \mathrm{diag}(\frac{\lambda}{\hat{w}})\right)\frac{\partial \hat{w}}{\partial B_i^k} &= B^T\frac{\partial \hat{r}}{\partial B_i^k} + \vec{\mathrm{e}}_i(\hat{r}_k - x_k) \\
\frac{\partial \hat{w}}{\partial B_i^k} &= -\left(B^T\frac{\partial \hat{r}}{\partial \hat{w}} + \mathrm{diag}(\frac{\lambda}{\hat{w}})\right)^{-1}\left(B^T\frac{\partial \hat{r}}{\partial B_i^k} + \vec{\mathrm{e}}_i(\hat{r}_k - x_k)\right)
\end{aligned} \tag{12}$$

This general form can be used with many loss/transfer function pairs by substituting the appropriate partial derivatives. The partial derivatives for the transfer transfer functions used in this paper are listed in the table below.

Transfer Function	$r = f(Bw)$	$\frac{\partial r}{\partial w}$	$\frac{\partial r}{\partial B_i^k}$
Linear	Bw	B	$\vec{\mathrm{e}}_k w_i$
Normalized Exponential	$\frac{e^{Bw}}{\sum_j e^{Bjw}}$	$(\mathrm{diag}(r) - rr^T)B$	$(\vec{\mathrm{e}}_k - r)w_i r_k$

D Text Application Details

Minimizing KL-divergence loss between the empirical probability distribution (type) of the document given in the input vector x and the reconstructed type after applying the normalized exponential transfer function $r = f(B, w) = \frac{e^{Bw}}{\sum e^{Bw}}$ (14) is equivalent to maximizing the "constrained poisson distribution" used to model documents in [1] , where N is the number of words in the document (13).

$$P(X|Nr) = \prod_i \frac{e^{-Nr_i}(Nr_i)^{X_i}}{X_i!} \tag{13}$$

$$\begin{aligned}
-log(P(X|Nr) &= \sum_i Nr_i - X_i \log Nr_i + \log\left(X_i!\right) = N\sum_i r_i - \bar{x}_i \log r_i + C \\
\mathrm{KL}(\bar{x}\|r) &= \sum_i r_i - \bar{x}_i \log r_i + \bar{x}_i \log \bar{x}_i - x_i = \sum_i r_i - \bar{x}_i \log r_i + C
\end{aligned} \tag{14}$$

Since our reconstruction and transfer functions are matched, the sparse coding update equations follow from substituting the normalized exponential transfer function into the general equations given in the implementation section.

E Music Genre Classification

A music genre classification task was also used for L_1 sparse coding in [2], and consists of 15, 60-second musical clips from each of 17 different genres. Following their practice, each song was divided into 50ms snippets, and the magnitude of the spectrogram for each snippet were used as input examples. For reconstruction, squared loss was used with a linear transfer function, and a maxent classifier was used for classification. The first 10 genres were used as unlabeled data to learn B via sparse coding, 10 clips from each of the other 7 genres were used as training data and 5 clips were used as testing data. As shown in Table 1, KL-regularization improved classification performance on the same basis over L_1-regularization.

Training Set Size	PCA	L1	KL
200	32%	31%	33%
2000	43%	43%	45%
10000	48%	49%	50%

Table 1: Classification Accuracy on a 7-way music genre classification task is increased by using the KL prior for sparse approximation.

References

[1] R. Salakhutdinov and G. Hinton, "Semantic hashing," in *SIGIR workshop on Information Retrieval and applications of Graphical Models*, 2007.

[2] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *ICML '07: Proceedings of the 24th international conference on Machine learning*, 2007.